Summary of Public Feedback Received on the 2010 Demonstration Data Product - Demographic and Housing Characteristics File

(2022-03-16)

Summary

The Census Bureau received 18 comments during the public comment period for the first round of **2010 demonstration data for the Demographic and Housing Characteristics File (DHC) (v. 2022-03-16)**. Of those 18 comments, 4 were unrelated to the demonstration data product (e.g., asking questions about release dates or reporting errors with the file). These were out of scope for the analysis. The remaining 14 comments were included in the analysis, with 11 providing feedback related to accuracy. All comments were provided through the 2020DAS@census.gov email address. See a full text compilation of in-scope comments received in Appendix B.

Many of the comments acknowledged improvements and fitness of use, but there were still concerns. On the person file, a common concern was accuracy of age data for geographies with smaller populations, smaller population groups, and off-spine geographies. On the unit file, multiple people expressed concern about the accuracy of household type, tenure, and household size. Several people expressed concerns about the number of inconsistencies between the person and unit file tables. They emphasized these inconsistencies would raise questions about the credibility and usability of the data. One person reported biases in the results that likely reveal an issue that will need more attention to fully understand.

Multiple people mentioned that 30 days did not allow for adequate review, and the Census Bureau would likely receive fewer comments due to the length of the review period.

Background

This document provides a summary of the comments received during the first round of 2010 demonstration data for the DHC.[1] The DHC person summary files and detailed summary metrics for the person and unit tables were released on March 16, 2022. The DHC unit summary files were released on March 29 and then rereleased on April 14 due to an error being identified. The public feedback due date was subsequently extended to May 16.

[1] The DHC demonstration data are consistent with the production settings demonstration data for the Redistricting Data (Public Law 94-171), and the 2020 DHC is expected to be consistent with the 2020 Redistricting data. Therefore, occupancy status, total population, race, Hispanic origin, people under and over 18 years, and major group quarters type is not expected to change.

The comments received during this period represent a cross-section of data users, including federal, state, regional, and local governments, academic institutions, private companies, and nonprofit organizations (Table 1). This was substantially fewer than the 400+ comments received on the 2020 Census Data Product Planning Crosswalk with a DHC comment period during September and October of 2021. The Crosswalk comment period sought feedback on the tables and geographic level needed by data users, including use cases. Feedback received on the Crosswalk helped inform the proposed design of the DHC. This more recent comment period (March 16 to May 16) focused on the accuracy achieved by first round DHC demonstration data.

Table 1. Comments by Type of Affiliation

Type of Affiliation	**Count**
Federal Government	1
State Government	3
Regional	1
Local Government	2
Academic Research	2
Private	4
Non-profit	1

Of the comments received, ten were concerned about accuracy, four about geography, two about public trust, one about equity, one about bias, and one about privacy (Table 2).[2] These concerns are not mutually exclusive as many people had multiple concerns. Some comments provided only general statements about the importance of accuracy for their uses of the data while others provided more in-depth analysis of differences between the published 2010 Census data and the 2010 demonstration data. Because the Census Bureau continues development on the disclosure avoidance system, some accuracy and data availability concerns have been alleviated or resolved. For example, based on data user feedback from the Crosswalk, we already planned to add back into the DHC a number of tables that were requested during that comment period. We plan to include these in the second round of DHC demonstration data.

Table 2. Commentor Concerns

Concern	**Count**
Accuracy	10
Geography	4
Public Trust	2
Equity	1
Bias	1
Release date	1
Privacy	1

[2] Note, 11 comments were related to accuracy, but only 10 comments had a concern about accuracy.

The feedback received covered many topics and subtopics (Table 3), and some topics were accompanied by a concern for accuracy. Multiple commentors mentioned the same concerns. Areas where multiple people were concerned included the accuracy of counts by sex and age, household type, tenure, and household size data.

In addition, many commentors were concerned with inconsistencies between the person and unit file tables. The person file includes person-level characteristics (e.g., Hispanic origin and race). The unit file includes characteristics of the housing unit, householder, or household (e.g., tenure and household type). Disclosure avoidance is applied to the person and unit files independently with no mechanism to merge the files or maintain consistency.

Although many comments expressed concerns with the same topics, there was disagreement on an acceptable level of accuracy. However, several commentors did provide their own thresholds for fitness-for-use and outlier values. There were concerns about the high number of inconsistent results between the person and unit files. Commentors thought that the high number of inconsistent results would lead to the perception that the data are not usable. Similarly, some commentors were concerned that the existence of outliers would keep people from using the data out of fear that they may end up basing a decision on an outlier value.

Table 3. Comments by Topics and Subtopics

Topic	**Subtopic**	**Count**
Age and/or Sex		12
	Age and Sex by Race and Hispanic origin	5
	Children	3
	School aged children	2
	Ages 0-4	2
	Median age	2
	Single year of age	2
	Ages <18, 18-64, and 65+	1
	5-year age categories	1
	Older age (70+ years)	1
	Working aged adults	1
	Age-adjusted ratios	1
	Age-specific rates	1
	Transgender	1
Person and Unit File Inconsistencies		6
	Persons per household	4
	Headship rates	3
Household Type		4
	Household Type by Presence of Children	1
Tenure		3
	Tenure by Race and Hispanic Origin	3

	Tenure by Age of Householder	1
Household Size		2
Total Population		3
Households or Housing Units		2
Race		1
Detailed Race and Hispanic Origin		1
Relationship		1
Noninstitutional Group Quarters		1
Age for Household Population		1
Families		1
Income		1

Geography

Commentors referenced a range of geographic levels in their analyses (Table 4). Many conducted their analysis at the tract or county level. Others expressed concern about the accuracy of off-spine geographies, including places and school districts.[3] Several mentioned the use of block-level data. While they acknowledged the Census Bureau's cautions about the use of unaggregated block data, they warned that many data users would continue to rely on block-level data regardless. In addition, they documented the impact of unreliable block-level data. For example, the Centers for Disease Control (CDC) commented that it would not be able to accurately characterize risks or identify and target vulnerable populations using block-level data. They said block-level maps would be unreliable regardless of any aggregation of the block-level data.

Table 4. Comments by Geography

Geography	**Count**
Block	4
Block Group	1
Tract	5
County	4
Place	4
School Districts	2
Off-spine	2
Urban/Rural	1

It was noted that block-level data often revealed inconsistencies due to the application of protections to the person and unit files independently. One commentor pointed out that more than 163,000 blocks only contained children (i.e., people under 18 years), which was viewed as a large difference from the

[3] Off-spine geographies are those geographies that do not nest perfectly within the Census geographic hierarchy that includes the nation, regions, divisions, states, counties, tracts, block Groups, and blocks.

published 2010 Census data. Based on their analysis, only 82 blocks had this same result in the 2010 Census data.

Concern about the Accuracy of Age Data

The person file includes data on age, sex, race, Hispanic origin, relationship to the householder, and group quarters. The primary concern for the person file was the accuracy of age data; all but one of the comments discussed age data. Commentors mentioned specific ages and subtopics for sex and age (Table 3). It should be noted that the second round of 2010 DHC demonstration data will provide additional age detail at the tract level.

Each commentor provided their own threshold for what would impact data fitness-for-use. For example, some commentors considered differences of 5 percent and 10 percent as large errors and differences of 25 percent as extreme errors. One commentor was concerned with error larger than 3 percent, and another commentor acknowledged it was challenging to identify which differences were impactful. Taken together, the comments provide a rough gauge for what is considered acceptable.

We note that average percent differences need to be considered with caution. Cells with small numeric differences can skew the average percent difference, creating the perception that many cells exhibit a high percent difference. Many commentors considered or accounted for this when providing their conclusions.

The analysis provided by most commentors focused heavily on the age data. It was recognized that age data are used in applications with many downstream impacts, such as population estimates that are used as controls for various surveys and denominators in rates. With the estimates traditionally relying on census data as its base, inaccuracies with the age data will be carried through the decade as each age cohort is aged forward each year.

Although most comments addressed accuracy of age data, there were a variety of subtopics. The most common subtopic was age data on children. These accuracy concerns included:

- Lower accuracy for off-spine geographies, namely school districts.
- High percent differences for smaller counties.
- Large differences when age groups were provided by race groups, especially smaller race groups.
- Accuracy for 4-year-olds.

In addition to data on children, commentors mentioned older ages, age-specific and age-adjusted rates, and person and unit file inconsistencies related to age. One commentor mentioned inaccuracies for older ages that we expect to be corrected in the second demonstration product. After recreating an earlier analysis with the new demonstration data, the CDC reported some improvements but also expressed concern about the usability of county and tract data for age-specific and age-adjusted rates. They examined the impact on rates developed for several programs at the county and tract levels. For each of the programs, they concluded there would be a “profound” and “significant” impact to the results. There were concerns about inconsistencies that might also be related to the accuracy of the age data. It was reported that in one city, there were more Native Hawaiian and Pacific Islander children than would be expected for the number of parents of the same race. These findings may be a result of the small cell sizes.

Many comments referenced geography in relation to age data. Larger differences for off-spine geographies were noted in a comparison between census designated places and villages. Concerns were expressed for less densely populated larger geographies where data users “expected better” accuracy even though the population groups were small. These findings could represent disparities in the usability of the data by age where populous areas have more usable data for lower geographic levels, such as tracts, and less-populated areas have higher percent differences for larger geographies, such as counties.

Counter to the concerns of some commentors, one analysis on median age found the age/sex distributions for the total population appear fit for use for on-spine geographies at the block group and above. However, they noted that differences for off-spine geographies (census designated places and school districts) are “much bigger.” One commentor from a larger city felt that the tract-level data for 5-year age groups for their city were fit-for-use: “Our finding is that while the demonstration files released in October 2019 and May 2020 could be wildly inaccurate, the latest release is indeed fit for use.”

Taken collectively, the analysis and comments relating to the accuracy of the age data acknowledge improvements and usability at some levels but continued concern about data for larger geographies, such as counties, and especially for off-spine geographies, including school districts. Commentor assessments of the data on age highlighted the lower accuracy for off-spine geographies. For tracts and even some higher geographic levels, the interpretation of the percent differences is that it would impact the usability of the data by age. Based on the comments received on the first DHC demonstration data, planning for schools and the calculation of rates related to public health would be impacted.

Concerns about Household Type, Household Size, and Tenure

The unit file provides the following information: tenure (rented vs owned), vacancy status, household size, household and family type, couple type, presence of people of specific ages, presence of own children, and householder characteristics (e.g., sex of householder).

Four people commented on household type, and one commented on household type and the presence of children specifically. Three people commented on household size. For example, one commented on lower accuracy for less common household types and smaller households: “Table cells for household type and household size that generally have lower counts (less common household types and household sizes) often have very large percentage errors (e.g., over 30 percent of tracts have more than 10 percent error), which severely limits the usability of these tables.”

Three people noted inaccuracies for tenure, specifically tenure by race and ethnicity. One person also noted a bias for tenure by age of householder. One commentor reported that the share of rented to total housing units was different for nearly all block groups and tracts in Virginia. When analyzing the same table for a subgroup of the population, there was a substantial amount of unusable results at the tract level. For example, Hispanic homeownership experienced a large change with 72 out of 520 (14 percent) tracts having differences equal to or greater than 10 percent. For one tract, the Hispanic homeownership value went from 21 in the published 2010 Census data to 38 in the differentially private protected data. They considered greater than 10 percent unusable and differences less than 5 percent acceptable. They stated: “This level of error makes this data very inaccurate and unusable for government policy-making and service planning efforts.”

Concerns about Person and Unit File Inconsistencies

Some data inconsistencies exist because disclosure protections are applied to the person and unit files independently; these inconsistencies are a feature of the disclosure avoidance system. While commentors generally understood that these inconsistencies could not be eliminated, almost half expressed concern about the number of them: “Of further concern are the number of blocks that have fatal inconsistencies which would invalidate any attempt to use the data to build either a custom geography (say a special district) or in tracking disasters, or public health issues.”

One commentor provided an accounting of these inconsistencies. These include:

- More households than household population.
- Household population without households.
- Householders not equal to households.
- Household population under 18 less than the number of households with children under 18.
- Not enough population to match a population calculated from household size.
- More householders of a certain age than population of that age.

Differences between counts of those living alone in the person and unit files were also reported. This situation is different than the other inconsistencies in that the person and unit tables include an explicit measure of the same characteristic. The DHC tables provide data on those living alone from a relationship table (P19) and a household type table (P16). The relationship table is created from the person file, and the household type table is created from the unit file. Thus, the living alone counts from table P19 do not match the same counts for table P16.

Commentors noted that “block-level data is full of inconsistencies,” but these inconsistencies were most concerning at the county and tract level. Tables provided by a commentor showed that when the same characteristic can be obtained from the person and unit files, the inconsistencies between the two files are pervasive impacting almost every geographic unit across all geographies. For example, the number of householders compared to the number of households. Inconsistencies that reflect an impossible result, such as a household population that is smaller than the number of households are not as pervasive but increase at the frequency of the population cell size decreases.

Concerns about Bias

One commentor identified a number of biases in unit file tables. An example of an observed bias was "areas with mostly rental occupied houses have a positive bias for households with children, whereas areas with mostly owner-occupied houses have a negative bias for households with children." It will take additional analysis to better understand these biases.

Other Topics

Some topics received less attention. Only one person commented on relationship to householder. Changes to the relationship categories from 2010 to 2020 may have limited the comments on accuracy. The demonstration data reflected the 2020 relationship categories, which did not exist in the published 2010 tables, resulting in issues of comparability. Strategies may need to be developed to obtain more feedback on the accuracy of the relationship to householder data for the second round DHC demonstration data. Although we received comments on race and Hispanic origin, they were in relation to other variables (e.g., tenure by race and ethnicity). We likely did not receive comments on race and Hispanic origin because those tables were released as part of the redistricting product. Because we expect consistency between the redistricting data and DHC, the accuracy for those tables has already been determined. We did receive one comment on the geographies for detailed race and ethnicity, but this comment applies to different data products (i.e., the Detailed Demographic and Housing Characteristics file [Detailed DHC-A] and Detailed DHC-B).

One commentor noted the importance of protecting privacy and conducted a simulation to examine the potential for identifying transgender children from two decades of census records. This was an innovative analysis that provided an example of how unprotected tabulations could be used to provide person-level information from the census. It highlighted the importance of protecting confidentiality and recommended the use of the TopDown Algorithm.

Some commentors were able to provide specific impacts if the data were released as-is. Cited impacts included funding for a variety of programs and policies, city and neighborhood planning (e.g., school-related needs), population projections, informed decision making (e.g., business planning and advertising), health disparities (e.g., prevalence and progression of cause-specific mortality), and small area estimation techniques. In addition, multiple people noted the potential impact on the Census Bureau, and census data losing public trust and credibility. Appendix A includes quotations regarding impacts.

Recommendations

Table 5 summarizes commentor recommendations. Most commentors recommended increased accuracy for age with some identifying specific topics (e.g., children) or geographies (e.g., school districts). In addition, many commentors recommended reducing the person and unit file inconsistencies. One supported the implementation of differential privacy via the TopDown Algorithm. There were recommendations to keep block-level data, add tables, and add geographies to tables already proposed for inclusion in the DHC. These comments echoed similar concerns when the Census Bureau collected public feedback on the Crosswalk. Two comments asked for tables with important use cases that were in 2010 Summary File 1 but were not included in the initial design of the DHC. Based on Crosswalk feedback, these tables are planned for inclusion in the second round DHC demonstration data product.

Table 5. Commentor Recommendations

Recommendation	**Count**
Increase accuracy for age	7
Reduce person and unit file inconsistencies	6
Increase accuracy for household type and/or household size	2
Increase accuracy for tenure	2
Increase accuracy for a specific geography	2
Keep block-level data	2
Add a table and/or geography	2
Increase accuracy for group quarters	1
Use TopDown Algorithm (TDA)	1

Conclusion

Overall, the commentors called for increased accuracy for age, household type, household size, and tenure—both alone and when crossed by race and ethnicity. Accuracy for what were considered key single years of age for children and school districts was a specific age-related concern. Commentors also expressed a concern about the person and unit file inconsistencies. While inconsistent results are accepted as a feature of the disclosure avoidance system, commentors expressed concern about the frequency of these inconsistencies. These comments suggest we need to improve the accuracy of the age data and unit file more generally. Greater accuracy across the unit file would not only improve the accuracy of the unit file tables, it likely would also reduce the number of person and unit inconsistencies.

Appendix A: Quotations Regarding Impacts
Appendix B: Full Text Compilation of Feedback Received

Appendix A
Quotations Regarding Impacts

"Federal and state funding received by schools and for educational planning... $39 billion of federal funds were distributed by the U.S. Department of Education to states and localities in FY 2017 based on census-derived data. Table 2 shows programs run by the U.S. Department of Education that distribute federal funds to state and localities based on census-derived data... 316 federal programs that use census-derived data to distribute about $1.5 trillion to states and localities in Fiscal Year 2017. About two-thirds of the 315 programs use substate data which underscores the important of small area census data. When one is talking about billions of dollars, a small percent error can translate into a large dollar amount."

"Population projections are often used to plan for expanding (or reducing) school facilities, staff, and other school-related needs. Current and projected demographic data are often used to construct attendance boundaries to keep classrooms from becoming overcrowded. Constructing attendance boundaries often include sensitivity to racial composition, so small area demographics by race are important."

"Federal agencies use this data to allocate federal funds directly to localities, such as Local Education Authorities, especially for special education and early intervention services, and for the expansion of Head Start, because those funding formulas include data on the number of young children. We note however that Title 1 funds are based on the number of children age 5 to 17."

"…impair the ability of businesses to make informed decisions, with potentially negative consequences for communities.

Values that even casual data users can identify as being incorrect lead to a lack of confidence in census bureau data and in our data by association."

"These data are another essential input for the preparation of population projections at the neighborhood-level using the cohort component model. Lacking such inputs, New York City will lose the precision we rely upon to direct billions of dollars in resources – resources directed towards a host of endeavors, from new school construction to the siting of our elder care facilities, essential elements for planning the future of our city.

In New York City, it is not enough to know, for example, that the Asian population has decreased in Manhattan's Chinatown. We must disentangle subgroup information by race, distinguishing whether it was the Chinese or Vietnamese population that declined in this example, so that we can properly allocate resources for services that our residents require."

"Block-level data are used by CDC for emergency response purposes to do environmental assessments when working with communities near environmental sites, and other analyses. CDC will not be able to

accurately characterize risks and identify/target vulnerable populations using block-level data. Block-level maps will be unreliable, regardless of any aggregation of the block-level data.

When age-sex data are used as progression ratios, large errors in one age group will corrupt the quality of older age groups as the model works its way forward. For this reason, we also need precise data for all cells in the age-sex table. Precision in these older age groups is especially important not only for our projections modelling, but also for other state departments (Health, Pollution Control, etc.) which track and compare the prevalence and progression of cause-specific mortality. These departments also track racial and ethnic disparities in disease prevalence."

"The shift of Housing units in Places between Occupied and Vacant is troubling as it directly affects popular small area estimation techniques such as the Housing Unit Method.

Of further concern are the number of blocks that have fatal inconsistencies which would invalidate any attempt to use the data to build either a custom geography (say a special district) or in tracking disasters, or public health issues."

Appendix B
Full Text Compilation of Feedback Received

2010 Demonstration Data Demographic and Housing Characteristics File (DHC) v. 2022-03-16

Round 1 Feedback

1. David Edmondson, Transportation Planner

As a transportation planner in a small city, I need access to solid population density data at the block-level with some minimum of fidelity:

- Population
- Homes

While good but not necessary, I also use the following at a block level, where available:

- School-aged children
- Working-age adults

Further, to ensure representation on committees and to receive feedback, I need at least tract-level – block and block-group is preferred – data on race, income, and age.

My look at the privacy-protective noise shows me that this will interfere with my ability to do my job effectively.

Please ensure that the population and housing data are accurate, that tract data is accurate, and that privacy is ensured through other means, such as confidence intervals, rather than fake noise "data" that would interfere with my analyses.

2. David Howell, Alaska Department of Labor and Workforce Development

Single year of age data at the county level still shows signs of clumping\binning.

For smaller county level geographies the single year of age data is quite spiky. This is smoothed out when summing to five-year age groups, or really just taking the average of two sequential years but single year data is still very important. I'm not sure if this issue is due to the DP model itself or from post processing but perhaps certain ages could be focused on to be more accurate. An example of an important age to know the population of is 4 year olds. School districts struggle the most with projecting the Kindergarten class size so it would be very good to know about how many kids will be entering the school district for the first time next year.

Ideally all of the ages would be more accurate but I'm not sure that this is possible with the privacy budget and time constraints on releasing the 2020 DHC data.

Minority age/race/sex data in small county level data also clumping/binning.

When the county level of data is examined by age/race/sex there is still evidence of people being lumped into some age groups and not others creating very spiky data, in some cases. In other cases the minority populations look fine but with the 2020 data release I will have no way of knowing which are accurate and which are not. Often times the actual age/race/sex data looks a bit odd due to the population being so small and there would be no way of knowing if this is the actual case or a result of the DP adjustment.

Another issue in the data are age structures that don't make sense. An example of this would be Native Hawaiian/Pacific Islander females in Juneau, AK. There are a large number of children belonging to the NHOPI Alone group but not enough mothers to have had this many children.

Again I'm not exactly sure what can be done to help here as I know these are the people that Title XIII was designed to protect and that the epsilon only allows so much accuracy. Having accurate data for these populations is imperative for the calculation of age/race/sex specific rates that can aid in better health in the future and the Census is the only place to get this data for small populations.

Place level age/sex data is unpredictable based on total population.

Alaska has many small places and many of those are CDPs. When looking at age/sex data at the place level it is hard to tell what the population cutoff should be for fitness of use. In general places with a total population of 1,000+ look pretty good when broken down by age/sex but this is not always the case.

For places under 500 it is definitely a mixed bag of accuracy. Some look great while others are totally unusable. Because of the sporadic nature of these smaller place's data it will likely not be possible to use any of the age/sex data from 2020 because we will have no way of knowing what is accurate and what is not. Age structures in places under 500 can be quite spiky in general so it's not possible to tell what's real and what is a result of the privacy adjustments.

Is there a plan to offer some type of guidance on what population level is needed to use the age/sex data at various geographic levels?

Persons per household.

We don't use much housing data in our estimates program but at a quick glance it looks like the number of households didn't change too much at the place level, I did not look at smaller geographies for this analysis. One issue that is still coming up though is a PPH of less than 1. There are five places in Alaska with a PPH below 1, the places are very small with less than 20 people but I still think it would be good to prioritize fixing this issue. It makes the data look quite bad and would be a very simple coding fix in the post processing. I'm sure there would be trickle down impacts but having impossible results in the data at the place level is not good for data quality metrics.

Thanks again for allowing outside organizations review the data!

3. Bill O'Hare, O'Hare Data and Demographic Services, LLC

Analysis of Census Bureau’s March 2022 Differential Privacy Demonstration Product: Implications for Data on Young Children

By

Dr. William P. O’Hare

Analysis of Census Bureau’s March 2022 Differential Privacy Demonstration Product: Implications for Data on Young children

By

Dr. William P. O’Hare

Executive Summary

The U.S. Census Bureau is using a new method called differential privacy (DP) to help protect confidentiality and privacy of respondents in the 2020 Census. This paper provides some information on how the use of DP in 2020 Census is likely to impact the accuracy of data for young children (population ages 0 to 4). The study is based on analysis of the most recent DP Demonstration Product released by the Census Bureau on March 16. 2022. The DP Demonstration Product issued in March 2022 supersedes earlier DP Demonstration Products and focuses on data for the 2020 Census Demographic and Housing Characteristics (DHC) file. This file has most of the tables that were in Summary File 1 in the 2010 Census. The Demonstration Product released in March 2022 has data for population and housing units, but this analysis only examines data from the population file.

This paper presents analysis of the error introduced by DP by comparing the data as reported in the 2010 Census Summary File to the same data after the application of DP. According to the Census Bureau, the demonstration file released by the Census Bureau in March has been optimized for major use cases of the DHC tables.

Analysis presented in this paper found little impact of DP on data about young children for large (highly aggregated) geographic units like states or large counties. However, the story is different for smaller geographic units. Many smaller areas have high levels of error in their data on young children after DP is applied. For example, the count of young children would exhibit absolute error of 5 percent or more in about 27 percent of Unified School Districts after DP is applied. The data also show that 69 percent of Unified School Districts had absolute numeric errors of 5 or more young children after DP is applied.

Errors of the magnitude shown above could have important implications for federal and state funding received by schools and for educational planning. Errors of this magnitude might impact formula funding that is based on Census-derived data and some schools will get less than they deserve.

Bigger absolute error percentages are evident for Hispanic, Black, and Asian young children in Unified School Districts. The mean absolute percent error for Non-Hispanic White young children was 5 percent compared to 27 percent of Hispanic young children, 34 percent for Black young children, and 42 percent for Asian young children. Differential accuracy among race and Hispanic Origin groups raises questions of data equity after DP is applied.

I also examined the accuracy/errors for the single year age 4 child population and found errors for single year of age are particularly large. I found 57 percent of Unified School Districts had absolute percent errors of 5 percent or more for children age 4, and 66 percent had absolute numeric errors of 5 or more children age 4.

Analysis also shows that 39 percent of Places (cities, village, and towns) had absolute percent errors of 5 percent or more for age 0 to 4, and 46 percent of Places had absolute numeric errors of 5 or more young children.

After the injection of DP in the 2010 Census data included in the March 2022 Census Bureau Demonstration Product, there were over 163,000 blocks nationwide that had population ages 0 to 17, but no population ages 18 or over. This result has two important implications, First, blocks with children and no adults is a highly implausible situation and the large number of such blocks may undermine confidence in the overall Census results. Second, these implausible results are likely due to young children being separated from their parents in 2020 Census DHC processing with DP. This separation of children and parent in data processing is an ongoing concern for data on young children and the production of future tables for children. This issue is particularly important in introducing DP into the American Community Survey, which is a key source of child well-being measures (O'Hare 2022b) To understand the well-being of children, it is critical to understand the situation of a child's parents or caretakers. - Moreover, if the same separation of children from their caregivers occurs in the application of DP to the American Community Survey, it will eliminate child poverty data which is based on household income. Child poverty data are the most important type of data on child well-being.

Based on the errors for young child population with the privacy parameters for DP used in the March 2022 DP Demonstration Product, and the lack of clarity about privacy protection from DP, I recommend the Census Bureau take steps to reduce the size of errors injected into the 2020 Census DHC file.

This paper is meant to provide stakeholders and child advocates with some fundamental information about the level of errors DP is likely to inject into the 2020 Census data for the population ages 0 to 4. There are a couple of reasons for sharing this information with child advocates now. The 2020 Census results for some localities may include situations where the number of young children reported looks suspect. It is important to make sure child advocates are aware of the potential impact of DP so they can explain odd child statistics to local leaders.

There is a second reason for sharing this information with state and local child advocates. The U.S. Census Bureau is looking for feedback on the use of DP in the 2020 Census. The Census Bureau is looking for cases where census data are used to make decisions. The Census Bureau is asking data users to examine the DP Demonstration Product to see if the error injected by DP make the data unfit for use. After reading this report, I hope you will convey your thoughts to the Census Bureau.

There is some latitude in how much error the Census Bureau will inject into the DHC files so feedback from census data users is important. If many users feel the current level of accuracy for data on young children in DP Demonstration Product is not accurate enough for some uses, there is a chance the Census Bureau could make the final data more accurate.

Stakeholders, child advocates, and data users should take advantage of this opportunity to communicate their thoughts to the Census Bureau before Census Bureau's Data Stewardship Advisory Committee makes a final decision on the privacy parameters before the DHC files are released in May of 2023. Comments on the implications of DP in the March 2022 Demonstration File are due Monday, May 16, 2022.

Comments and responses can be sent to 2020DAS@census.gov.

Analysis of Census Bureau's March 2022 Differential Privacy Demonstration Product: Implications for Data on Young children

By

Dr. William P. O'Hare

Introduction

The U.S. Census Bureau is using a new method called differential privacy (DP) to help protect confidentiality and privacy of Census respondents in releasing data from the 2020 Census.[1] This paper uses several measures to assess the accuracy of census data for young children after DP is applied. Young children are defined in this report as those ages 0 to 4. The analysis is based on the Demonstration Product data released on March 16, 2022, which is the most recent available from the Census Bureau.

In short, DP injects errors in the data provided by respondents to make it more difficult for someone to be identified in the Census records. Adding or subtracting random numbers to the census results makes it more difficult to identify data for specific respondents because the data in the published census results no longer match what respondents submitted. The U.S. Census Bureau (2020e) provides more information

[1] The terminology in this arena can be confusing. Differential Privacy is sometimes called "formal privacy." The system developed for the 2020 Census DHC file has also been called the Top Down Algorithm or TDA. Since the application of differential privacy occurs within the Census Bureau's Disclosure Avoidance Systems (DAS) that term has sometimes been used to describe the use of Differential Privacy. To avoid confusion, I use the term differential privacy (DP) here to distinguish the version of DAS that includes DP from other versions of DAS.

on the use of DP in the 2020 Census along with regular updates of their work (U.S. Census Bureau 2020c). In the fall of 2021, the Census Bureau released a primer on DP. (U.S. Census Bureau 2021d).

For an independent look at differential privacy see Boyd (2019) or Bouk and Boyd (2021). Hotz and Salvo (2020) offer a good review of DP early in the Census Bureau’s development. . A good overview of the evolution of the DP issue at the Census Bureau is provided by Boyd and Sarathy (2022). I think it is fair to say that the introduction of DP in the 2020 Census has become a very controversial issue. In their review of the development of the DP issue over the past few years, Boyd and Sarathy (2022, page 1) conclude, “When the U.S. Census Bureau announced its intention to modernize its disclosure avoidance procedures for the 2020 Census, it sparked a controversy that is still underway.”

One reason to focus on impact of DP on the population ages 0 to 4 is the high net undercount of that population in the Census. Results of the 2020 Census evaluation, using the Demographic Analysis method, shows a net undercount of 5.4 percent for young children which was much higher than any other age group (U.S. Census Bureau 2022c). Recent trends are also unsettling. From 1950 to 1980, the young children and adults had similar decade-to-decade improvement in terms of census coverage. However, after 1980 the trajectories were quite different. The coverage for adults continued to improve while the coverage of young children decreased dramatically (O’Hare 2022a).

There are a couple of perspectives one could take regarding the high net undercount of young children and DP. On one hand, since the 2020 Census data for young children already has more error than data for other age groups, perhaps the amount of error injected by DP should be limited for this group. It doesn't seem fair to inject more error into data for groups that already have a lot of error in their census data. On the other hand, one might think that since the 2020 Census data for young children already has a lot of error, the added error from DP will not make much difference.

I focus first on data accuracy for Unified School Districts because schools are the public institution most closely associated with the child population and schools use demographics in a variety of ways. I next look at data for Places. Places include big cities and small villages. They typically have policymaking authority, and they often provide programs for young children such as childcare or preschool programs.

Several issues regarding DP are addressed in the Discussion section included the high error rate for blocks, breaking the relationship between children and parents,questions of equity, and the extent to which DP contributes to the lack of public trust.

Background on Privacy in the Census

In every census, the U.S. Census Bureau faces a trade-off between privacy protection and accuracy. According to the U.S. Census Bureau (2020d),

> "One of the most important roles those national statistical offices (NSOs) play is to carry out a national population and housing census. In so doing, NSOs have two data stewardship mandates that can be in direct opposition. Good data stewardship involves both safeguarding the privacy of the respondents who have entrusted their information to the NSOs as well as disseminating accurate and useful census data to the public."

The problem that DP is designed to fix is complicated as is the implementation of DP. The passage below from the U.S. General Accountability Office (2020, page 14) is the best short description I have seen on this issue.

> “Differential privacy is a disclosure avoidance technique aimed at limiting statistical disclosure and controlling privacy risk. According to the Bureau, differential privacy provides a way for the Bureau to quantify the level of acceptable privacy risk and mitigate the risk that individuals can be reidentified using the Bureau’s data. Reidentification can occur when public data are linked to other external data sources. According to the Bureau, using differential privacy means that publicly available data will include some statistical noise, or data inaccuracies, to protect the privacy of individuals. Differential privacy provides algorithms that allow policy makers to decide the trade-offs between data accuracy and privacy. “

It is important to note that the U.S. Census Bureau has used methods to help avoid disclosure of individual census respondents for many decades. According to U.S. U.S. Census Bureau (2018) some method of disclosure avoidance has been used by the U.S. Census Bureau since 1970. The 2010 Census data include some changes to original responses to help avoid disclosure of information about individual respondents, largely using a method called swapping.

The application of Differential Privacy allows the Census Bureau to control the amount of error injected into the data which is largely controlled by something called “Epsilon.” A higher-level epsilon means less error and more risk of violating confidentiality and a lower epsilon means more error and less risk of violating confidentiality. In the latest material from the Census Bureau, Epsilon has been replaced with a term called Rho. It is my understanding Rho works the same way as Epsilon in that a higher value means more accuracy.

Measuring Accuracy

There is no consensus on exactly what measures should be used to assess the accuracy of DP-infused data, and there is no single benchmark to determine if DP-infused figures are "accurate enough for use." The U.S. Census Bureau (2020a) has suggested several measures of accuracy that could be used to evaluate the DP-infused data.

For simplicity I only look at a few key measures here, but I believe they provide sufficient information to reach some conclusions. The measures used here (mean absolute numeric error, mean absolute percent error, and large errors) are a subset of those discussed by Census Bureau. Like the Census Bureau's assessment of DP-infused data, I provide data for both absolute numerical errors and absolute percent errors because either can be important and using both perspectives provide a more complete picture of the error profiles for geographic units.

The DP demonstration file released by the Census Bureau on March 16, 2022, provides DP-infused data from the 2010 Census which can be compared to the 2010 Census data without DP to understand the likely impact DP has on data accuracy.

Errors are defined here as the difference between the data as originally reported in the 2010 Census Summary File and the same data after DP has been injected. The data from the Summary File is sometimes referred to as data without the application of DP in this report. Specifically, I subtract the value of the data without DP (Summary File) from corresponding DP-infused data to find the error. For percentages, the difference is divided by the data without DP (i.e., Summary File) value.

I include a measure the Census Bureau calls the Mean Absolute Error (I label this Mean Absolute Numerical Error in the tables to distinguish it from the Mean Absolute Percent Error) and I also include the Mean Absolute Percent Error.

An absolute error reflects the magnitude of the error regardless of direction. A geographic unit with an absolute error of 10 percent could be 10 percent too high or 10 percent too low. Absolute errors are used to make sure positive errors and negative errors do not cancel each other out and make it appear as if there are no errors.

Percent error reflects the size of the error relative to the size of the population. An error of a given magnitude (say 10 young children) may be trivial in large Places but very significant in smaller Places. For example, a numeric error of 10 young children in a school district of 1,000 young children is only a 1 percent error, but a numeric error of 10 young children in a school district of 100 is a 10 percent error.

In addition to measures of average error, I include analysis on the number and percent of geographic units that have relatively large errors. I use two sets of benchmarks to identify large errors: one for absolute numeric errors and one for absolute percent errors.

I believe the number and percent of large errors are likely to be the most important measures of accuracy in the 2020 Census. Large errors are likely to be a statistical problem and a public relationship problem for the Census Bureau, particularly if they are accompanied by large swings in funding that are not connected to real changes in population size. Data from the Census is often used to distribute federal and state dollars based on population (O'Hare 2020a: Reamer 2020). Large errors can result in implausible or impossible results. Such results are likely to cast suspicion on all the data from the Census Bureau and it is likely to undermine the confidence people have in all the census data.

Data Used in This Study

The Demonstration Product released in March 2022 reflects ongoing work at the Census Bureau. Starting in October 2019, the Census Bureau has released several Demonstration Products that reflect the injection of DP into 2010 Census data. The first official data from the 2020 Census with DP infused was the redistricting data file released by the Census Bureau in August 2021.

The data used in my analysis were originally provided by the Census Bureau. The IPUMS- NHGIS unit at the University of Minnesota processed the Census Bureau files and put the data into more user-friendly tables. I analyze the data produced by IPUMS-NHGIS unit which are available at https://nhgis.org/privacy-protected-demonstration-data

According to the IPUMS- NHGIS unit at the University of Minnesota, the privacy loss budget assigned to person-level and housing unit-level counts in the 2022-03-16 vintage file was 20.82 and 22.77, respectively. This contrasts to an Epsilon of 19.6 used for the 2020 Census redistricting files.

Geographic units where there were zero people ages 0 to 4 in either the 2010 data with DP or without DP, were removed from the file for analysis. Observations with zeros for key measures produce very unusual results. This analysis does not include data for Puerto Rico.

Results for Age 0 to 4 in Four Kinds of Geographic Units

Table 1 provides a few key accuracy measures for the population ages 0 to 4 for four kinds of geographic units. These units were selected because they all have significant policy-making power regarding programs for children.

The results shown in Table 1 indicate that DP is unlikely to have much of an impact on the young child data for states. The mean absolute numeric error for states for the population ages 0 to 4 is about 7 young children and the mean absolute percent error rounds to zero.

Also, DP is unlikely to have much impact on young child county data for most counties. The mean absolute numeric error for counties is about 8 young children and mean absolute percent error is 0.9.

However, of the 3,221 counties examined here 35 percent (1,140) had less than 1,000 children ages 0 to 4. For this subset of counties, DP may distort the data to a considerable degree. For the 302 counties with less than 5,000 people, the mean absolute percent error for ages 0 to 4 was 4.6 percent and the mean numeric error was 5.

Table 1 Key Statistics for Absolute Numeric and Absolute Percent Errors* for Children Ages 0 to 4 for Selected Geographic Units

	States	Counties or county equivalent	School Districts	Places
Number of Units in the Analysis	50	3,221	10,864	28,729
Mean Size of District (Children ages 0-4 based on Summary File)	39,873	6,342	1,880	546
Mean Absolute Numeric Error**	7	8	12	6
Mean Absolute Percent Error	rounds to zero	0.9	4.3	13.6
Percent of Units with Absolute Numeric Errors of 5 or more young children	58	62	69	46
Percent of Units with Absolut Percent Errors of 5% or more	0	3	27	39

Source: Author's analysis of Demonstration Product data released by the Census Bureau on March 16, 2022 after being prociess by IPUMS NHGIS at the University of Minnesota www.nhgis.org

Data in this table does not include Puerto Rico or geographic units with zero population age 0 to 4 in 2010 Summary File or DP-infused file.

* in this paper errors reflect the difference between the 2010 Census data without and with DP injected.

** The Census Bureau calls this measure Mean Absolute Error. I include the word "Numeric" to distinguish it from Mean Absolute Percent Error.

DC is not included in the state data but is included in the county data

The situation is different for Unified School Districts and Places (shown in Table 1), where DP is likely to cause larger distortions (percentagewise) for the young child population. The mean absolute numeric error for Unified School Districts is 12 young children and it is 6 young children for Places. The mean absolute percent error for United School Districts is 4.3 percent and it is 13.6 percent of Places.

Accuracy for Unified school Districts and Places are explored in more detail in the next two sections of this report.

Application of Differential Privacy to School District Data

The analysis first focuses on Unified School Districts since schools are the largest public institution focused on children. The Census Bureau reports there were 61.6 million children ages 3 to 17 enrolled in schools in 2019 (U.S. Census Bureau 2021a).

Schools often provide preschool programs for those under age 5. The Census Bureau shows there were over 5 million children enrolled in preschool in 2019, and more than half of all children age 3 and 4 are in preschool or nursey school (McElrath et al. 2022)

Reamer (2020) shows that $39 billion of federal funds were distributed by the U.S. Department of Education to states and localities in FY 2017 based on census-derived data. Table 2 shows programs run by the U.S. Department of Education that distribute federal funds to state and localities based on census-derived data. In addition, many other government programs also use census-derived data to distribute funds targeted to children.,

Overall, Reamer (2020) identified 316 federal programs that use census-derived data to distribute about $1.5 trillion to states and localities in Fiscal Year 2017. About two-thirds of the 315 programs use substate data which underscores the important of small area census data. . When one is talking about billions of dollars, a small percent error can translate into a large dollar amount. This is even more true when the funding allocation is based only on a particular age group.

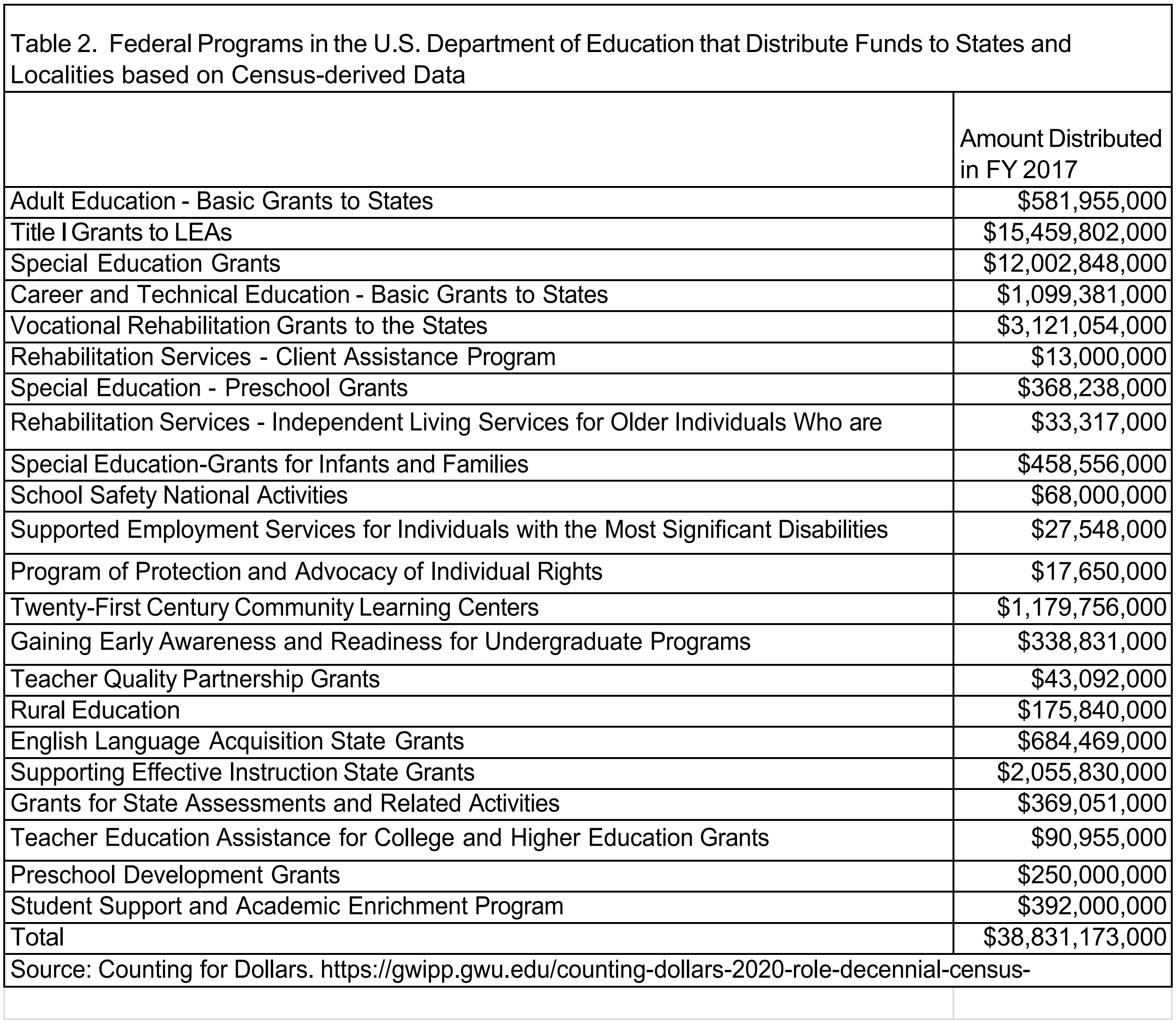


Table 2. Federal Programs in the U.S. Department of Education that Distribute Funds to States and Localities based on Census-derived Data

	Amount Distributed in FY 2017
Adult Education - Basic Grants to States	$581,955,000
Title I Grants to LEAs	$15,459,802,000
Special Education Grants	$12,002,848,000
Career and Technical Education - Basic Grants to States	$1,099,381,000
Vocational Rehabilitation Grants to the States	$3,121,054,000
Rehabilitation Services - Client Assistance Program	$13,000,000
Special Education - Preschool Grants	$368,238,000
Rehabilitation Services - Independent Living Services for Older Individuals Who are	$33,317,000
Special Education-Grants for Infants and Families	$458,556,000
School Safety National Activities	$68,000,000
Supported Employment Services for Individuals with the Most Significant Disabilities	$27,548,000
Program of Protection and Advocacy of Individual Rights	$17,650,000
Twenty-First Century Community Learning Centers	$1,179,756,000
Gaining Early Awareness and Readiness for Undergraduate Programs	$338,831,000
Teacher Quality Partnership Grants	$43,092,000
Rural Education	$175,840,000
English Language Acquisition State Grants	$684,469,000
Supporting Effective Instruction State Grants	$2,055,830,000
Grants for State Assessments and Related Activities	$369,051,000
Teacher Education Assistance for College and Higher Education Grants	$90,955,000
Preschool Development Grants	$250,000,000
Student Support and Academic Enrichment Program	$392,000,000
Total	$38,831,173,000

Source: Counting for Dollars. https://gwipp.gwu.edu/counting-dollars-2020-role-decennial-census-

It is also clear that census-related data are often used by states to distribute state government money, but as far as I can tell, there is no systematic data on how much money is distributed by states based on Census data (O’Hare 2020a).

At the Committee on National Statistics DP workshop held in December 2019 there were several presentations reflecting implications of DP-infused data for young children and school districts (Vink 2019; O’Hare 2019; Nagle and Kuhn 2019:). O’Hare (2021) focuses on the accuracy of population ages 0 to 17 for Unified School Districts based on data from the Census Bureau’s redistricting file. Note that some of these analyses are now outdated but may be useful for framing issues.

Demographic data are used for several important school district applications. Population projections are often used to plan for expanding (or reducing) school facilities, staff, and other school-related needs. Demographic projections are typically based on Decennial Census data.   Current and projected demographic data are often used to construct attendance boundaries to keep classrooms from becoming overcrowded. Such activities often require very small area data such as census blocks. Demographers who work extensively with school districts report that census blocks are a critical geographic unit for their work (Cropper et al. 2021). Constructing attendance boundaries often include sensitivity to racial composition, so small area demographics by race are important.

Many school districts are governed by school boards which are often elected from single member districts.   Such districts must meet the usual legal requirements of redistricting such as having districts with equal population size. Such redistricting must also meet the requirements of the Voting Rights Act, which means small area tabulations of population by race and Hispanic origin are important.

Once children get into the K-12 school system, school systems have pretty good data for forecasting the number of children to expect in each grade the following year. From that perspective it is the cohort age 0 to 4 that is the biggest unknown for school systems. Therefore, this is the most important age group for examining the amount of error injected by DP.

Districts where there was a zero for population age 0 to 4 in the DP or SF file were not included in the analysis. Also, recall Puerto Rico is not included.

DP has a bigger impact, percentage wise, in smaller populations and the majority of Unified School Districts are relatively small. Many of the 10,864 Unified School Districts are very small; 729 of the Unified School Districts had total population less than 1,000, and 3,875 districts had total population less than 5,000 in the 2010 Census. The translation of small numeric errors into large percent errors is also more apparent in looking at data for Hispanic, Black, and Asian groups within school districts because those are typically smaller population groups.

Table 3 shows several measures of accuracy/error for 10,864 Unified School Districts in the 2010 Census used in this analysis. The data are provided for all young children (all races) as well as for Non-Hispanic White Alone young children, Hispanic young children, Black Alone young children, and Asian Alone young children. For the remainder of this report when I use the term Black or Asian, it means Black alone or Asian alone. Other race groups were not examined here because the numbers were small, they were often highly clustered, and time was limited.

Data in Table 3 show the vast majority of Unified School Districts have at least one Black child, one Hispanic child, and one Asian child. But many districts have few young children of color. The average number of Hispanic young children in School Districts where there was at least one Hispanic was 524, for Blacks it was 396, and for Asians it was 151. These numbers are well below the overall average of 1,880 young children. The relatively small number of Black, Hispanic, and Asian young children in many districts results in these groups having larger absolute percent errors.

Table 3 shows the mean absolute numeric error for all young children (all races) in Unified School Districts is 12 young children. Data in Table 3 shows for all children, the mean absolute percent error as 4.3. But these measures mask big differences among race and ethnic groups

The mean absolute numeric errors for race and Hispanic Origin groups are smaller than for all children (10 for Non-Hispanic White Alone young children, 7 for Hispanic young children, 5 for Black young children, and 4 for Asian young children), On the other hand, mean absolute percent error was 4.3 percent for all children, 27 percent for Hispanics, 34 percent for Blacks young children, and 42 percent for Asian young children (see Table 3).

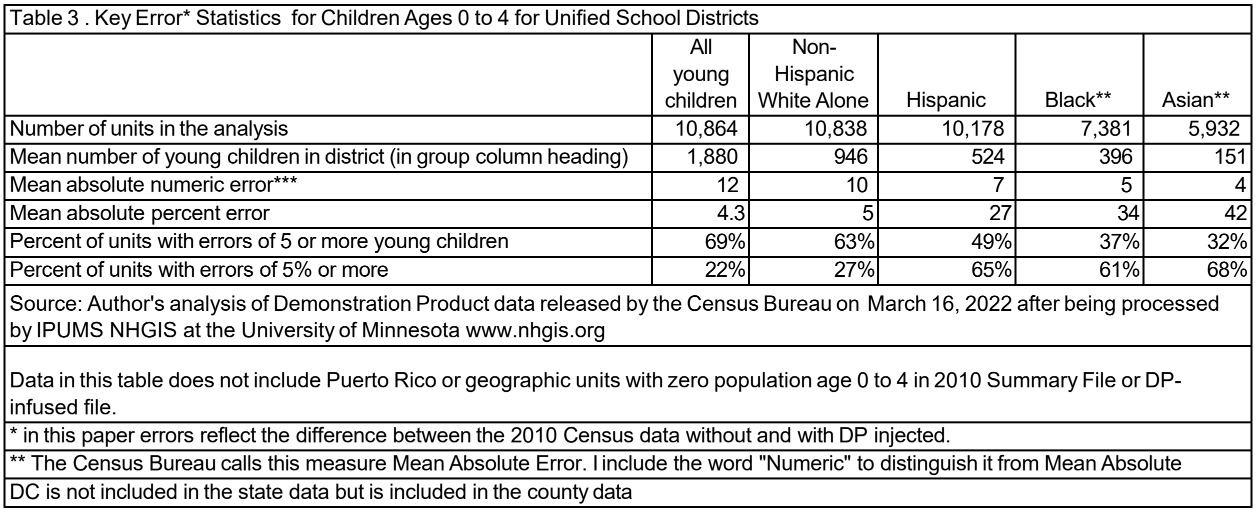


Table 3 . Key Error* Statistics for Children Ages 0 to 4 for Unified School Districts

	All young children	Non-Hispanic White Alone	Hispanic	Black**	Asian**
Number of units in the analysis	10,864	10,838	10,178	7,381	5,932
Mean number of young children in district (in group column heading)	1,880	946	524	396	151
Mean absolute numeric error***	12	10	7	5	4
Mean absolute percent error	4.3	5	27	34	42
Percent of units with errors of 5 or more young children	69%	63%	49%	37%	32%
Percent of units with errors of 5% or more	22%	27%	65%	61%	68%

Source: Author's analysis of Demonstration Product data released by the Census Bureau on March 16, 2022 after being processed by IPUMS NHGIS at the University of Minnesota www.nhgis.org

Data in this table does not include Puerto Rico or geographic units with zero population age 0 to 4 in 2010 Summary File or DP-infused file.

* in this paper errors reflect the difference between the 2010 Census data without and with DP injected.

** The Census Bureau calls this measure Mean Absolute Error. I include the word "Numeric" to distinguish it from Mean Absolute

DC is not included in the state data but is included in the county data

Recall that absolute errors reflect the magnitude of the error without regard to the direction of the error. Absolute errors are used so that positive and negative errors do not cancel each other out in constructing an average or mean.

Large Errors in Unified School Districts

Means or averages are helpful, but they do not reveal the full story. Large errors can be problematic even if the overall mean is relatively low. An examination of the distribution of Unified School Districts by error size can provide more information on the relative accuracy of the DP-infused data.

There is no consensus on what constitutes a large error and definitions probably vary with different applications. I show three benchmarks for large absolute percent errors. The 5 percent or more and 10 percent or more categories are used in several publications. I added the 25 percent plus category to look at the most extreme errors. Errors of 25 percent or more are likely to be very problematic. These thresholds are judgmental, but I think they provide a reasonable range of errors.

To be clear, the districts with more than 25 percent with large errors are also counted in the categories for more than 10 percent error and more than 5 percent error.

Distributions of absolute percent errors are shown in Figure 1 which shows that for all young children, 22 percent of districts had absolute percent errors of 5 percent or more for all children, compared to 27 percent of Non-Hispanic White Alone, 65 percent for Hispanic young children, 61 percent for Black young children, and 68 percent for Asian young children. Since minority groups are smaller in population size, it is not surprising that there are more extreme absolute percent errors.

Figure 1 also shows that for young children of color, absolute percent errors of 25 percent or more are not unusual.

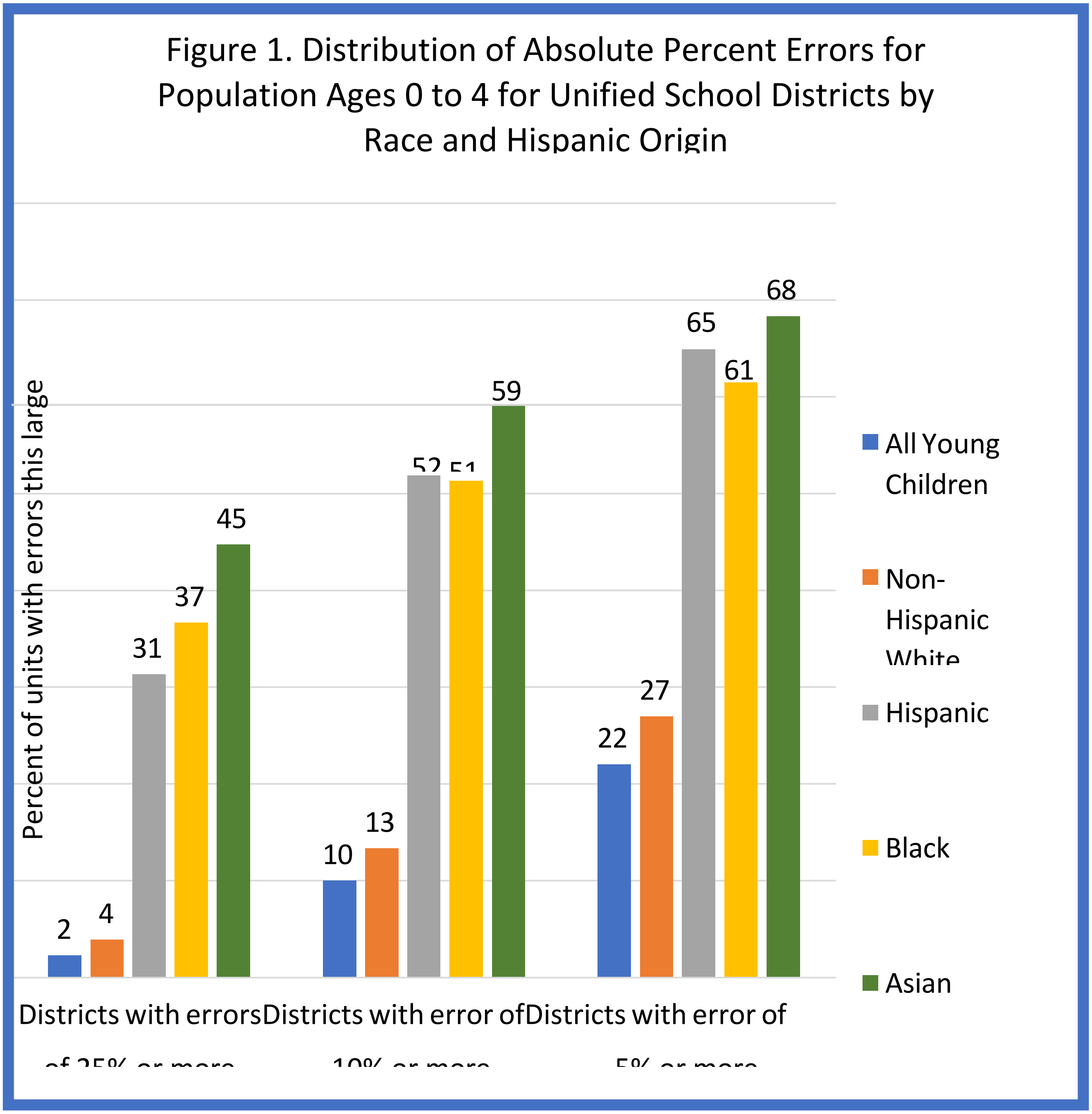


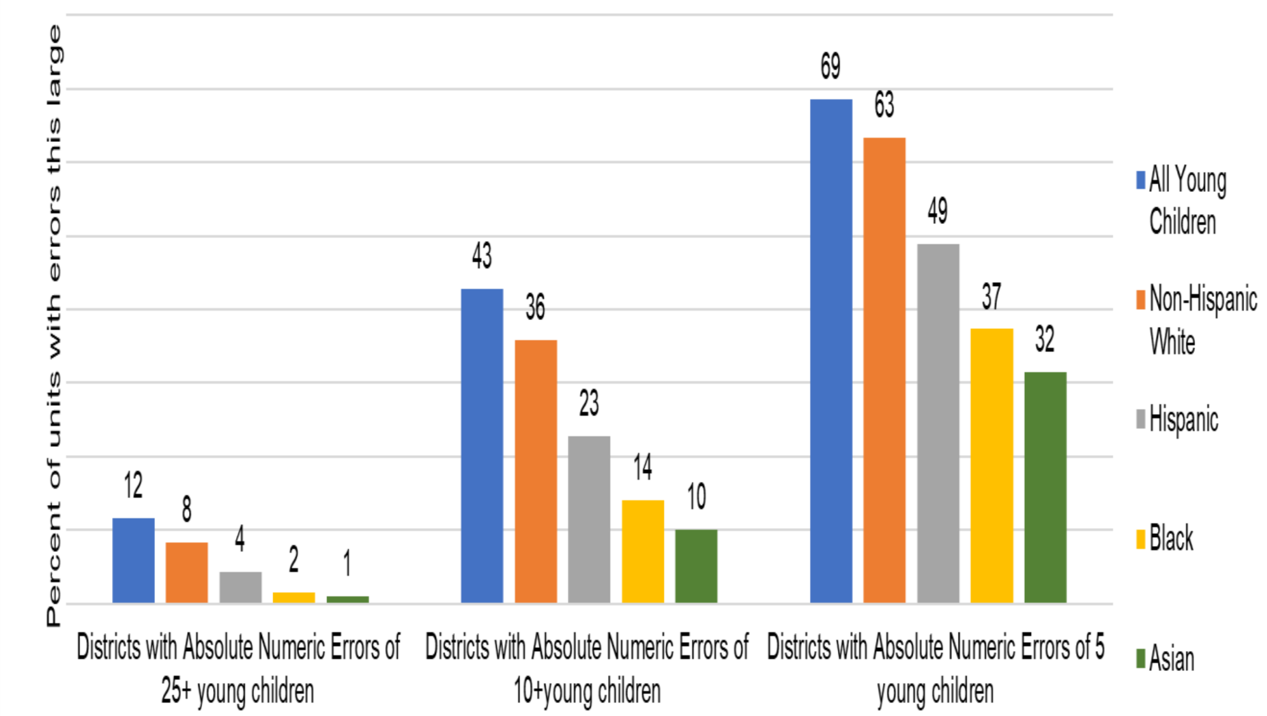
Figure 1. Distribution of Absolute Percent Errors for Population Ages 0 to 4 for Unified School Districts by Race and Hispanic Origin

I use three benchmarks for large absolute numeric errors. The 5 person and 10 person categories of errors have been used in other publications. I added the 25 persons plus category to look at the most extreme errors. Errors of 25 or more young children are likely to be very problematic.

Figure 2 shows 69 percent of the Unified School Districts had errors of 5 young children or more for young children of all races but the figures for minority groups are smaller: 49 percent for Hispanic young children, 37 percent for Black young children, and 32 percent for Asian young children.

In Figure 2, in each category of absolute numeric errors (5 young children, 10 young children, and 25 young children), there are many fewer districts that have this level of error for Hispanic, Black, and Asian young children than there are districts that have this level of error for all young children or Non-Hispanic White young children.

Figure 2. Distribution of Absolute Numeric Errors for Population Ages 0 to 4 for Unified School Districts by Race and Hispanic Origin



There are relatively few Unified School Districts with very large absolute numeric errors. Only 12 percent of Unified School Districts have errors of 25 young children or more, compared to 4 percent of Hispanic young children, 2 percent for Black young children, and 1 percent for Asian young children.

The national numbers shown above mask a lot of variation across states. Table 4 shows states ranked on two key measures of accuracy (mean absolute numeric error and mean absolute percent error) for Unified School Districts. The mean absolute numeric error for states ranges from a low of 0 for Hawaii (Hawaii only has one unified school district) to a high of 32 for Montana. The mean absolute percent error ranges from a low of 0 for Hawaii to a high of 6.7 percent in South Dakota.

Table 4 States Ranked by Mean Absolute Numeric Error and Absolute Percent Error for Ages 0 to 4 by Unified School Districts

Rank*		Average of absolute numerical error		Rank*	State	Average of absolute percent error
1	Montana	32		1	South Dakota	6.7
2	Maine	25		2	Nevada	6.4
3	North Dakota	24		3	New York	5.9
4	Washington	20		4	Oklahoma	5.7
5	Nebraska	19		5	New Hampshire	5.4
6	South Dakota	19		6	Iowa	5.3
7	Oklahoma	19		7	Texas	5.2
8	Oregon	18		8	North Dakota	5.1
9	Vermont	17		9	Alaska	5.1
10	Idaho	17		10	Wisconsin	4.8
11	Colorado	17		11	Montana	4.8
12	Texas	16		12	Arkansas	4.7
13	New Mexico	16		13	Colorado	4.6
14	Alaska	16		14	Ohio	4.4
15	Kansas	16		15	Illinois	4.3
16	Missouri	15		16	Oregon	4.3
17	Iowa	15		17	Nebraska	4.3
18	Wyoming	15		18	Michigan	4.3
19	Arkansas	13		19	Kansas	4.2
20	New Hampshire	12		20	Pennsylvania	4.2
21	New York	12		21	Missouri	4.2
22	Michigan	12		22	Minnesota	4.1
23	Minnesota	12		23	Idaho	3.9
24	Illinois	11		24	New Mexico	3.9
25	Wisconsin	11		25	Washington	3.9
26	Ohio	10		26	Connecticut	3.9
27	Nevada	10		27	Arizona	3.9
28	Arizona	9		28	Tennessee	3.7
29	Indiana	8		29	Utah	3.7
30	Mississippi	7		30	Wyoming	3.6
31	California	7		31	Mississippi	3.5
32	Kentucky	6		32	West Virginia	3.4
33	Delaware	6		33	Massachusetts	3.3
34	Pennsylvania	5		34	Indiana	3.1
35	Tennessee	5		35	California	3.0
36	South Carolina	5		36	Georgia	3.0
37	Utah	5		37	Virginia	3.0
38	New Jersey	5		38	Maine	2.9
39	Virginia	4		39	Kentucky	2.9
40	Alabama	4		40	New Jersey	2.8
41	Massachusetts	4		41	Maryland	2.6
42	Georgia	4		42	Vermont	2.4
43	West Virginia	4		43	Alabama	2.4
44	Rhode Island	4		44	Florida	2.4
45	Connecticut	4		45	North Carolina	2.2
46	North Carolina	4		46	Rhode Island	2.1
47	Louisiana	3		47	South Carolina	1.8
48	Florida	2		48	Louisiana	1.8
49	Maryland	2		49	Delaware	1.1
50	Hawaii	0		50	Hawaii	0.0
U.S. Average		12		U.S. Average		4.3

Source: Author's analysis of Demonstration Product released by the Census Bureau on March 16, 2022 afterprocessing by IPUMS NHGIS at the University of Minnesota www.nhgis.org

*Ranks are based on unrounded data.

Analysis for Age 4

In the Demonstration Product released in March 2022, the Census Bureau provided data by single year of age and sex for the population under age 20. I analyze this data for age 4 for Unified School Districts. I selected age 4 because that is often used by school systems to predict the number of kindergarteners to expect in the following school next year. I do not see any reason why the metrics for age 4 would be much different than the metrics for any other single year of age.

Table 5 provides the key metrics for the comparison of age 4 in Unified School Districts in the 2010 Census file with and without DP. Districts with no people age 4 in the DP or SF file were not used in the analysis. The mean absolute numeric error was 11 and the mean absolute percent error was 11 percent for age 4

A big share of Unified School Districts had large errors in both numeric and percentages terms. Two-thirds (66 percent) of Unified School System had absolute numeric errors of 5 or more children and 57 percent of Unified School Districts had absolute percent errors of 5 percent or more for children age 4.

With errors of this magnitude for single year of age, one has to wonder if this data is worth producing, particularly for small districts. It is not clear how users are supposed to manage data with this degree of uncertainty.

Table 5. Unified School District Error* Metrics for Age 4

Number of Units in Analysis	10,424
Mean number of 4 year old's in Summary File	394
Mean Absolute Numeric Error	11
Mean Absolute Percent Error	11
Percent of units with Absolute Numeric Error 5+ children age 4	66
Percent of units with Absolute Percent Error 5%+	57

Source: Author's analysis of Demonstration Product released by the Census Bureau on March 16, 2022 after processing by IPUMS NHGIS at the, University of Minnesota www.nhgis.org

* In this paper, errors reflect the difference between the 2010 Census data without and with DP injected.

Data in this table does not include Puerto Rico or geographic units with zero population age 0 to 4 in 2010 Summery File or DP-Infused file.

Data for Places

Census Places are geographic units used by the U.S. Census Bureau to publish data. They range from Places with millions of people such as Los Angeles and New York City, to the smallest villages and towns.

Places include both incorporated Places and Census Designated Places (CDPs). There are a little more than 29,000 Places for which the infusion of DP data was produced in the March 16, 2022 (DP Demonstration Product) and most of them (over 19,000) are Incorporated Places rather than Census Designated Places (CDPs). Incorporated Places are legally bounded entities such as cities, boroughs, towns, or villages (names may vary depending on the state). Census Designated Places (CDPs) are statistical entities used in the Census. They are unincorporated communities where there is a concentration of population, housing, and commercial structures and they are identifiable by name. There are nearly 10,000 CDPs for 2010 Census data.

Cities, villages, and towns might want to know about the number of young children in their area for things like planning youth activities, child facilities, and day care centers. The preschool-age population is also useful for forecasting future school enrollments.

The mean absolute numeric error for Places was 6 and the mean absolute percent error was 13.6 percent. The high percent error is not surprising because many of these Places are small. There were 1,422 Places where the number of young children was less than 100, and 9,012 Places where the number of young children was less than 500, based on the 2010 Summary File.

Figure 3 shows the distribution of Places by absolute percent error using the same thresholds used for Unified School Districts. The data in Figure 3 shows that almost half (46 percent) of Places had absolute percent errors of 5 percent or more for the young child population and 15 percent had absolute percent errors of 25 percent or more. Since Places are generally smaller (in population size) than Unified School Districts, it is not surprising that the percentages are larger for Places than for Unified School Districts.

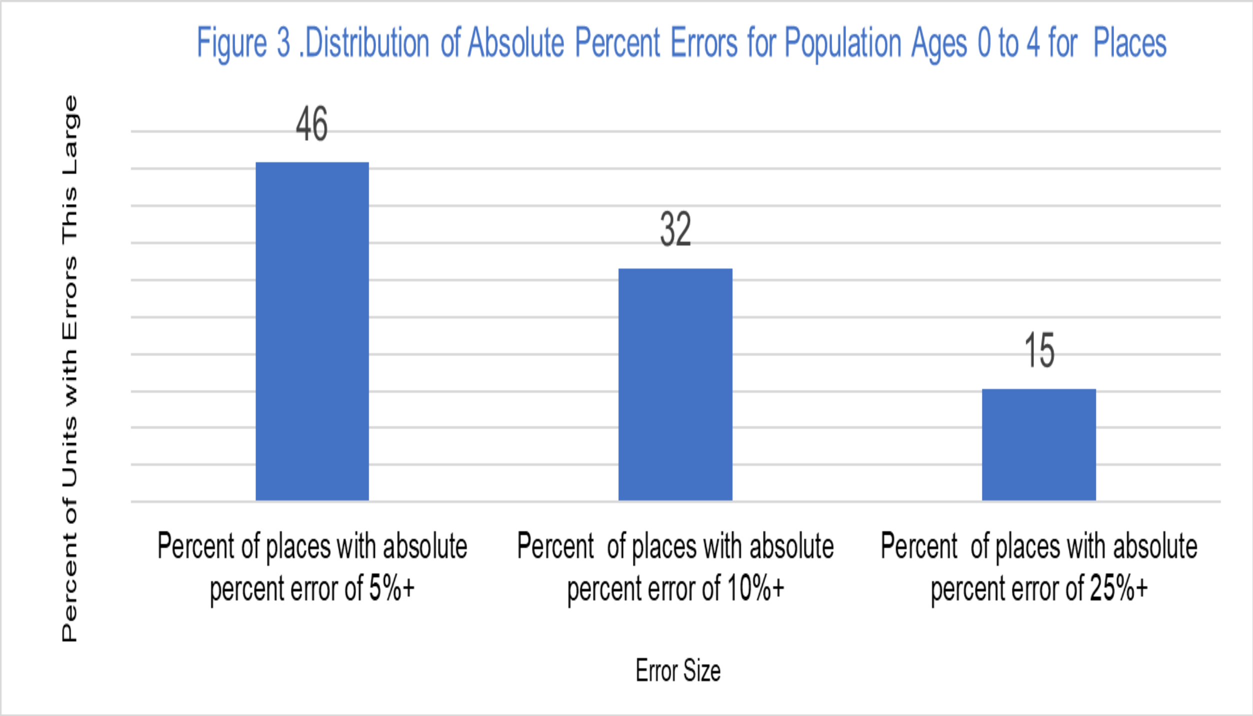


Figure 4 show the distribution of Places by absolute numeric errors using the same categories as Figure 2. Data show 39 percent of the Places had absolute numeric errors of 5 or more young children, and only 2 percent had absolute percent errors of 25 or more young children.

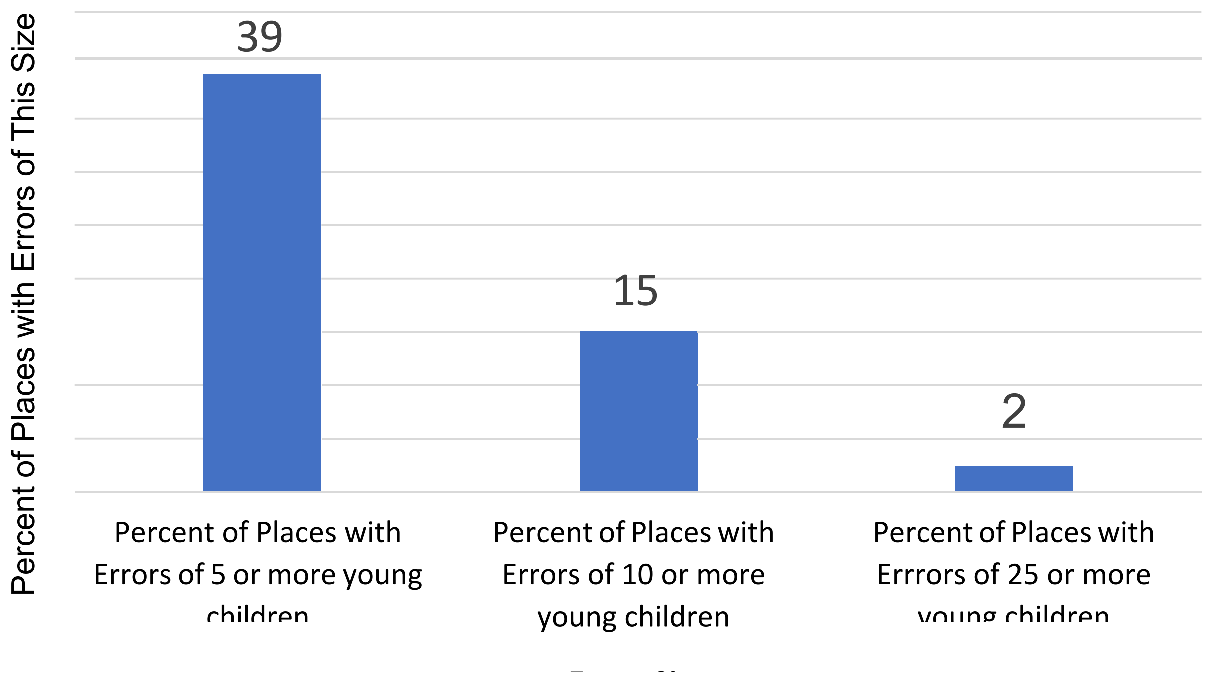


Figure 4. Distribution of Places by Absolute Numeric Errors for Population Ages 0 to 4

Table 6 shows states ranked on the percent of places in a state with absolute percent errors of 5 percent or more. Data for errors of 10 percent or more and 25 percent or more are also provided in the Table 6.

There is a lot of variation across the states. For example, 68 percent of the Places in Vermont had absolute percent errors of 5 percent or more, compared to 27 percent in New Jersey.

Table 6 States Ranked By Percent of Places is State with Absolute Percent Errors of 5 or more for Population Ages 0 to 4

		Number of Places in State	Percent Distribution Within State		
Rank		State Total	Absolute Percent errors of 5+	Absolute Percent errors of 10+	Absolute Percent errors of 25+
1	Vermont	117	68	50	23
2	New Mexico	414	65	53	30
3	New Hampshire	95	65	42	18
4	Montana	339	64	52	34
5	Alaska	307	62	48	26
6	North Dakota	342	62	50	31
7	Wyoming	182	61	51	26
8	South Dakota	352	61	48	27
9	Nebraska	543	59	44	26
10	Oklahoma	700	58	43	20
11	West Virginia	393	58	42	20
12	Arizona	426	58	46	23
13	Kansas	647	56	42	23
14	Maine	130	53	35	14
15	Iowa	977	51	35	17
16	Rhode Island	34	50	32	18
17	Nevada	122	49	41	26
18	Missouri	996	49	34	17
19	Arkansas	530	49	33	10
20	Virginia	587	48	33	18
21	Pennsylvania	1,741	48	34	15
22	Colorado	429	48	35	19
23	Minnesota	887	47	32	15
24	Kentucky	519	45	28	13
25	New York	1,178	45	28	11
26	North Carolina	731	44	30	13
27	Idaho	217	44	31	12
28	Connecticut	142	44	29	10
29	Washington	608	44	30	14
30	Maryland	502	43	32	21
31	Texas	1,714	43	29	16
32	Oregon	368	43	32	17
33	Wisconsin	760	43	28	13
34	Ohio	1,197	42	28	13
35	Michigan	686	42	28	11
36	Alabama	571	42	28	14
37	Indiana	675	42	27	12
38	South Carolina	392	40	26	12
39	Utah	319	40	27	9
40	Louisiana	469	40	24	9
41	California	1,458	39	28	16
42	Delaware	75	39	21	12
43	Illinois	1,359	39	24	9
44	Georgia	618	38	23	8
45	Massachusetts	242	37	20	6
46	Tennessee	427	35	22	7
47	Mississippi	362	35	20	8
48	Hawaii	150	34	23	7
49	Florida	909	31	18	8
50	New Jersey	536	27	19	9
	U.S Total	28,474	46	32	15

Source: Author's analysis of Demonstration Product released by the Census Bureau on March 16, 2022 after processing by IPUMS NHGIS at the, University of Minnesota www.nhgis.org

* In this paper, errors reflect the difference between the 2010 Census data without and with DP injected.

Data in this table does not include Puerto Rico or geographic units with zero population age 0 to 4 in 2010 Summery File or DP-Infused file.

there was 1 place with no state code

Table 7 shows states ranked on the percent of places in the state with absolute numeric errors of 5 or more young children. Data for 10 percent or more and 25 percent or more are also shown in the table. There is a lot of variation among the states. For example, 74 percent of places in Rhode Island have absoltue numeric errors of 5 or more young children comapred to 15 percent of North Dakota.

Table 7. States Ranked by Percent of of Places with Absolute Numeric Error of 5 or more for Population Ages 0 to 4

			Percent of Places with Errors This Large		
Rank	Row Labels	Number of Places in the State	errors of 5 or more young children	erros of 10 or more young children	errors of 25 or more young children
1	Rhode Island	34	74	32	3
2	Maine	130	67	35	2
3	Hawaii	150	66	38	8
4	Connecticut	142	64	43	11
5	Massachusetts	242	61	35	7
6	California	1453	60	32	7
7	New Hampshire	95	58	28	5
8	Maryland	501	57	28	7
9	Virginia	587	56	30	9
10	Florida	908	53	28	5
11	New York	1178	53	23	3
12	Arizona	422	51	25	4
13	Washington	605	51	25	5
14	New Jersey	535	49	23	4
15	Nevada	121	48	23	4
16	Texas	1708	46	17	2
17	Vermont	117	45	21	2
18	Utah	317	45	19	3
19	Michigan	686	45	15	2
20	New Mexico	412	44	17	3
21	South Carolina	392	44	20	4
22	Delaware	75	43	20	5
23	Louisiana	470	43	20	4
24	Colorado	426	41	16	2
25	North Carolina	730	40	16	2
26	Oregon	365	40	14	2
27	Pennsylvania	1739	39	16	3
28	Ohio	1196	38	13	3
29	Georgia	617	37	14	2
30	West Virginia	392	36	11	1
31	Tennessee	427	36	14	1
32	Montana	339	33	11	0
33	Alabama	571	33	11	1
34	Oklahoma	698	32	7	0
35	Indiana	674	32	11	1
36	Kentucky	518	32	8	0
37	Wyoming	181	32	10	2
38	Wisconsin	758	31	9	1
39	Illinois	1356	31	9	1
40	Alaska	303	31	10	2
41	Mississippi	362	29	11	0
42	Idaho	216	26	8	0
43	Minnesota	883	26	5	0
44	Arkansas	529	24	5	0
45	Kansas	641	24	4	0
46	Missouri	987	24	5	1
47	South Dakota	350	20	7	1
48	Iowa	968	18	1	0
49	Nebraska	536	15	2	0
50	North Dakota	332	15	3	0
	U.S Total	28,476	39	15	2

Source: Author's analysis of Demonstration Product released by the Census Bureau on March 16, 2022 after processing by IPUMS NHGIS at the, University of Minnesota www.nhgis.org

* In this paper, errors reflect the difference between the 2010 Census data without and with DP injected.

Data in this table does not include Puerto Rico or geographic units with zero population age 0 to 4 in 2010 Summery File or DP-Infused file.

there was 1 place with no state code

Discussion

It is clear the introduction of DP into the 2020 Census has caused a lot of controversy. I have been following the U.S. Census since 1970, and I do not remember any issue that has caused as much discussion, concern, and debate among data users as the decision to implement DP in the 2020 Census.

Below I review a couple of issues regarding DP that were not addressed in my analysis but may impact stakeholders view of DP

Block-Level Data

Blocks are the smallest geographic unit used in the Census and there are about 8 million blocks in the 2020 Census but only about 6 million are occupied. The average block has a total population of about 41 people and about 3 young children. The small population size of blocks makes them susceptible to large percent errors when random numbers are injected with DP.

The availability of errors at the Census block level makes DP different that normal assessment of Census accuracy. Assessment of Census accuracy using the two standard Census Bureau methods (Demographic Analysis and Post-Enumeration Survey) is only available at the national level as this report is being written (state data will come out soon from PES but there will be no substate data from these methods). But the DP Demonstration Product allows one to look at errors for all levels of Census geography down to the census block level.

There are two broad perspectives on the error DP injects into census blocks. One perspective is that data for census blocks are among the most important data supplied by the Decennial Census, and they need to be as accurate as possible. One of the primary purposes of the Decennial Census is to provide comparable population figures for small areas across the country. To the best of my knowledge, there is no other data source that provides demographic data for all the blocks in the country other than the Decennial Census. Consequently, census accuracy for blocks is especially important. O'Hara (2022) makes a strong case for why block level data are important in terms of creating special or custom districts. The need for such data is often not apparent until well after the Census data has been collected and reported.

Another perspective holds that blocks are typically aggregated into larger units like congressional districts, cities, and counties and in those aggregations the random error injected into blocks cancel each other out and produce relatively accurate data for larger units. From this perspective, errors at the block level are not so important.

Regarding the usability of block level data, the Census Bureau (Devine 2022, slide 17) recently stated, "Block-level data are fit-for-use when aggregated into geographically contiguous larger entities. They are not intended to be fit-for-use as a unit of analysis." It seems likely the high level of inaccuracy for census blocks based on analysis of the DP Demonstration Product influenced the issuance of this statement.

I do not think there is any dispute that the error injected by DP for blocks produces a relatively high absolute percent error and that these errors typically cancel each other out when blocks are aggregated into larger areas. Because the error is random, the amount of error does not become cumulative. It is an open question about how important census block level data are for making decisions.

One problem with use of DP for small areas is the implausible or impossible results. I did not have the computer power to examine blocks for age 0 to 4 in the March 2022 Demonstration Product, but Census Bureau 2022 data show heavy distortions at the block level . For example, more than 163,000 blocks have children (population age 0 to 17) but no adults (population age 18 and over) after DP is applied compared to just 82 such blocks before DP was applied (U.S. Census Bureau 2022_). Many such cases are highly unlikely and raise questions about who these children are living with if there are no adults in their household. The Census Bureau (2022d) offers several other examples of implausible or impossible results in the data after DP is applied

It is not clear to me exactly what statistical problems might be caused by these results, but they undermine the veracity of the census data broadly. A high number of improbable results is identified as a problem of "legitimacy" rather than statistical accuracy by Hogan (2021) and is likely to undermine the confidence the public has in the Census results. When data users see highly implausible results like the large number of blocks with children and no adults, they often wonder what other errors are in the data that are not so apparent.

Despite the statement by the Census Bureau and misgivings among some demographers about the quality of census block data, many data users routinely use the block level data, either because they do not realize the level of potential errors, or because it is the best (or only) data they have at that level.

The data indicate the average percent errors for census blocks is relatively high but does not address how often block-level data are used in decision making. Readers may have their own answer to that question.

Breaking the Link Between Child and Parents

The production of many blocks where there are children, but no adults may be related to the link between children and adults in a household that is broken when 2020 DAS with DP was applied to the DHC file. DP is administered to children and parents independently, so it may eliminate the adults in a household that has children by randomly subtracting data from the number of adults. If the processing retained the link between young children and their parents in a household, it is doubtful that there would be such a high number of blocks with children and no adults.

This statistical disconnection of children and parents is an on-going concern and is likely to have important impacts in later Census products which have more detailed data on young children.[2] For example the connection between children and parents is critical for a lot of data from the American Community Survey. Child poverty is probably the single most important measure of child well-being and determining poverty status requires linking a child to the income of the adults in the households.

The Census Bureau says it will use a different method of DP in the Detailed Demographic and Housing File which will retain the connection between children and

[2] It is my understanding that the use of DP does not necessarily require the disconnect between children and parents in a household. The break between children and parents in the redistricting file and the DHC is a result of the particular DP-related processing chosen by the Census Burau.

parents. Hopefully, that will alleviate our concerns. But data that links children and adults in the Detailed Demographic and Housing file will not be available until late 2023 or 2024. That is getting very close to the date (2025) the Census Bureau said it might start applying DP to the American Community Survey (ACS) Translating the application of DP from the Census to the ACS, is likely to be a difficult process.

Accuracy and Equity

The focus of this report is on census accuracy, but the differential accuracy raises the issue of equity. Equity in terms of data provision has become a more visible aspect of data collection and reporting in the federal government recently (White House Equitable Data working Group 2022). According to the U.S. Census Bureau (2021e, pages 1) " The Census Bureau has an ongoing commitment to producing data that depict an accurate portrait of America, including its underserved communities." Data equity has become a part of broader equity questions. This suggests all results should be examined through the lens of equitable data.

In terms of equity, Figure 1 shows substantial differential accuracy for Unified School Districts by race and ethnicity after DP infused. For Hispanic young children, the mean absolute percent error was 27, for Black young children the mean absolute percent error was 34, and for Asian young children was 42, compared to 5 for Non-Hispanic white children. What does this say about the equity of using the DP method? There is already differential accuracy in census results before DP is applied but it may be the case that DP exacerbates such inequities. Is if fair to inject as much error for groups that already have a lot of error in census data as for those groups that do not have much error? Did the Census Bureau examine equity concerns when they decided to use DP in the 2020 Census?

Selection of a DAS and Public Trust

Disclosure avoidance is not just a statistical issue and examining it only from a statistical perspective may be problematic. Another dimension for assessing alternative DAS methods is the extent to which a given DAS method undermines public trust in the Census data and the Census Bureau itself. There has been a great deal of concern about the erosion of public trust in the Census Bureau recently. According to the National Academy of Sciences, Engineering and Medicine panel assessing the 2020 Census (2022, page 6),

"We are very concerned, based on presentations to the panel and our knowledge of reactions to previous demonstration data, that the Census Bureau's adoption of differential privacy-based disclosure avoidance has increased the level of public mistrust in the 2020 Census and the Census Bureau itself."

In their review of the impact DP has had on the Census Bureau credibility and trust among data users, Boyd and Sarathy (2022, page 1) state, "We argue that rebuilding trust will require more than technical repairs or improved communication: it will require reconstructing what we identify as a "statistical imaginary."

Summary

This report provides information on accuracy of DP-infused data and provides a profile of the likely errors for young children that will be seen in data for in the 2020

Census if the Census Bureau uses the privacy protection parameters reflected in the March 2022 Demonstration Product.

It is important to note that the analysis provided in this paper is just a sample of analyses that could be done. But I believe the data analyzed in this study a relatively good sample of the broader implications of using a DAS method with DP in the DHC with the privacy protection parameters used in this Demonstration Product.

But there are many other data factoids that could have been produced to shed light on the implications of DP. For example, the Census Bureau shows the mean absolute percent error for foster children at the county level is 122 percent and at the Incorporated Place level is 96 percent after DP has been applied. Foster children are very vulnerable population, and this level of error is disturbing.

The question that is not addressed in the previous sections is whether the level of error reflected in this analysis would make 2020 Census for data on young children "unfit for use." Each person will probably have a different answer to how much error in census data for young children is too much error.

Like all disclosure avoidance systems, the use of DP involves a trade-off between privacy protection and census accuracy. There have always been errors in the Census data, but in the 2020 Census, the Census Bureau is trying to decide how much additional error to add to the data in order to enhance privacy protection. By setting privacy parameters, the Census Bureau has control over the level of accuracy and level of privacy protection in the 2020 Census.

Given this balancing act, it would be useful to have more information about two aspects of DP: 1)metrics on privacy protection and 2)information on potential harm of re-identification that DP is designed to protect against. It would be helpful if we could compare the metrics of accuracy to metrics of privacy protection in the March 2022 Demonstration Product. I see many measures of accuracy based on the Demonstration Product. However, I don't see any privacy protection metrics produced by the Census Bureau nor do I see a way to explore the privacy protection aspect with the Demonstration Product. It seems the balance of accuracy and privacy protection is the key reason for using a given disclosure avoidance systems but without metrics for privacy protection I am not sure how to do that.

In assessing the tradeoff between privacy and accuracy it is not clear exactly what harm might be done by a re-identification of someone based on Census data. My recommendation below might be different if lower privacy protection meant hundreds of innocent people would go to jail versus few people getting annoying phone calls. But I have not seen any evidence on this question from the Census Bureau. When I have asked experts about the level of privacy protection afforded by an Epsilon of 19.6 in the redistricting data in terms I can understand it seems like I always get a variation of "it depends." But no metrics.

On the other hand, the problems that are likely to be caused by inaccurate census data on young children are clearer to me. The data in this paper, and many other analyses, provide a rich set of metrics showing the magnitude of error DP injects into Census data and I can envision problems such errors might cause.

When the number of young children in a school district is under-reported by 5 or 10 percent, that could have big implications for their funding and when the number of young children in a community is off by 10 percent or more, that could impact planning in ways that waste taxpayer money and undermine quality education for young children. If the number of young children reported in the Census for a Unified School District is 10 percent too low, it may not automatically translate into 10 percent less money for that jurisdiction. But there is a strong link between underreporting the number of young children and the loss of money in a general sense.

In addition to the money distributed on the basis of census-derived data, Census data are used for many decisions in the public and private sector. The more errors there are in the data, the less likely those decisions will be correct ones.

Given the level of errors in Unified School Districts and Places using the privacy protection level in the most recent DP Demonstration Product, and the lack of clear evidence or measurements about the level or impact of privacy loss, I recommend that the Census Bureau increase the level of accuracy used in the DHC to provide more accuracy small area data for young children.

Author Note

It should be noted that this analysis is not as full and complete as it should be because time did not allow such an analysis. The Census Bureau released the latest DP Demonstration Product on March 16, 2022 but had to re-release it on Apirl 14 because of mistake They request responses by May 16th

Since stakeholders need time to read and absorb this paper it needed to be available well before May 16th. If there had been more time for analysis there is a lot more that could have been done. The data used here could be developed to provide a more granular picture of DP's impact. For example, one could calculate the measures shown here for all counties or all Places within a state, or one could develop the measures for all census tracts within a county.

If more time had been available, it would have been useful to explore data for race and Hispanic groups more thoroughly. Also, it would have been useful to examine accuracy measures for geographic units of different population sizes. If I had more time, I would have used race alone or in combination rather than race alone. There is a good deal more that could be done to provide state-specific data.

It is unfortunate that the time limitations mean the Census Bureau will not receive the quality of feedback they seek.

References

Bouk, D. and Boyd, D. (2021*). Democracy's Data Infrastructure.; The technologies of the U.S. Census*. https://knightcolumbia.org/content/democracys-data-infrastructure

Boyd. D. (2019). "Balancing Data Utility and Confidentiality on the 2020 US Census," Data and Society, https://datasociety.net/library/balancing-data-utility-and- -confidentiality-in-the-2020-us-census/ .

Boyd, D. and Sarathy, J (2022) "Differential Perspectives: Epistemic Disconnects Surrounding the US Census Burau's Use of Differential Privacy," *Harvard Data Science Review* (forthcoming)

Committee on National Statistics (2019). "Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations," presentations are available at https://www.nationalacademies.org/event/12-11-2019/workshop-on-2020-census-data-products-data-needs-and-privacy-considerations .

Cropper, M. McKibben and Stojakovic, Z. (2021). The Importance of Small Area Census Data for School Demographics, Count all Kid website https://ednote.ecs.org/counting-all-kids-how-the-census-impacts-education/

Hogan, H. (2021). "The History of Assessing Census Quality, Presentation at 2021 Population of Association of America Conference, May 5, 2021.

Hotz, J. and Salvo J. (2020). Addressing the Use of Differential Privacy for the 2020 Census: Summary of What We Learned from the CNSTAT Workshop. https://www.apdu.org/2020/02/28/apdu-member-post-assessing-the-use-of-differential-privacy-for-the-2020-census-summary-of-what-we-learned-from-the-cnstat-workshop/ ,

McElrath, K. Bauman, K., and Schmidt, E (2022) Preschool Enrollment in the United states,: 2055 to 2019," U.S. Census Bureau https://www.census.gov/content/dam/Census/newsroom/press-kits/2021/paa/paa-2021-presentation-preschool-enrollment-in-the-united-states.pdf

Nagle, N. and Kuhn, T. (2019). "Implications for School Enrollment Statistics." https://www.nationalacademies.org/event/12-11-2019/workshop-on-2020-census-data-products-data-needs-and-privacy-considerations.

National Academy of Sciences, Engineering and Medicine, (2022). *Understanding the Quality of the 2020 Census, Interim Report* , Washington Dc. The National Academy Press, https://nap.nationalacademies.org/catalog/26529/understanding-the-quality-of- the-2020-census-interim-report

O'Hara, A. (2022) presentation at Analysis of Census Noise Measurements Workshop, April 28-29, Rutgers University.

O’Hare, W.P. (2019). Assessing 2010 Census Data with Differential Privacy for Young Children,” https://www.nationalacademies.org/event/12-11-2019/workshop-on-2020-census-data-products-data-needs-and-privacy-considerations .

O’Hare W. P. (2020a). “Many States Use Decennial Census Data to Distribute State Money, The Census Project Website https://thecensusproject.org/2020/01/09/many-states-use-decennial-census-data-to-distribute-state-money/

O’Hare, W.P (2020b). “Implications of Differential Privacy for Reported Data on Young children in the 2020 U.S. Census,” Posted on Count All KIDS Website Implications-of-Differential-Privacy-for-kids-11-17-2020-FINAL-00000003.pdf (myftpupload.com) .

O’Hare, W.P. (2021). “Analysis of Census Bureau’s August 2021 Differential Privacy Demonstration Product: Implications for Data on Children,” Count All Kids website November *https://countallkids.org/resources/analysis-of-census-bureaus-august-2021-differential-privacy-demonstration-product-implications-for-data-on-children/*

O’Hare, W. P. (2022a). “New Census Bureau Data Show Young Children Have a High Net Undercount in the 2020 Census, “ Posted on Count All Kids website , March, https://countallkids.org/resources/new-census-bureau-data-show-young-children-have-a-high-net-undercount-in-the-2020-census/

O’Hare , W.P (2022b). U.se of the American Community Survey Data by State Child Advocacy Organizations? Count All Kids website, https://countallkids.org/resources/use-of-the-american-community-survey-data-by-state-child-advocacy-organizations/

Reamer, A. (2020). Counting for Dollars, George Washington University https://gwipp.gwu.edu/counting-dollars-2020-role-decennial-census-geographic-distribution-federal-funds .+

U.S. Census Bureau (2018), “Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing,” THE RESEARCH AND METHODOLOGY DIRECTORATE, Mc Kenna, L. U.S. Census Bureau, Washington DC., https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/Disclosure%20Avoidance%20for%20the%201970-2010%20Censuses.pdf .

U.S. Census Bureau (2019). “2010 Demonstration Data Products,” U.S. Census Bureau, Washington DC., October, https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html .

U.S. Census Bureau (2020a). 2020 Census Disclosure Avoidance Improvement Metrics, U.S Census Bureau, Washington DC., March 18, https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-03-18-2020-census-da-improvement-metrics.pdf?# .

U.S. Census Bureau (2020b), "2020 Census Data Products and the Disclosure Avoidance System", Hawes M. and Garfinkel. S. L., Planned presentation at the Census Scientific Advisory Committee meeting, March 26.

U.S. Census Bureau (2020c) DAS Updates, U.S Census Bureau, Hawes M. June 1 Washington DC., https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-06-01-das-updates.pdf?# .

U.S. Census Bureau (2020d). "Disclosure Avoidance and the Census," Select Topics in International Censuses, U.S. Census Bureau, October 2020. https://www.census.gov/library/working-papers/2020/demo/disclos-avoid-census.html .

U.S. Census Bureau (2020e). "Disclosure Avoidance and the 2020 Census, U.S. Census Bureau," Washington DC., Accessed November 2, https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html .

U.S. Census Bureau (2020f). "Error Discovered in PPM," U.S. Census Bureau, Washington DC. https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html .

U.S. Census Bureau (2020g), "2020 Disclosure Avoidance System Updates," U.S. Census Bureau, Washington DC., https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html .

U.S. Census Bureau (2021a) School Enrollment in the United States: October 2019 - PowerPoint Presentation (census.gov Detailed Tables, School Enrollment in the United States: October 2019 - Detailed Table 1, FEBRUARY 02, 2021.

U.S. Census Bureau (2021b) Developing the DAS: Demonstration Data and Progress Metric, https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-development.html .

U.S. Census Bureau (2021c). "Differential Privacy 101." Webinar May 4, 2021, Michael Hawes. https://www.census.gov/data/academy/webinars/2021/disclosure-avoidance-series/differential-privacy-101.html

U.S. Census Bureau (2021d). " Disclosure Avoidance for the 2020 Census: An Introduction," U.S. Census Bureau, Washington, DC. November https://www.census.gov/library/publications/2021/decennial/2020-census-disclosure-avoidance-handbook.html

U.S. Census Burearu (2021e) “Advancing Equity with Census Bureau Data.” Census Bureau Blog, November 2, 2021, Ron Jarmin , Acting Director Advancing Equity With Census Bureau Data

U.S. Census Bureau (2021f), “Disclosure Avoidance for the 2020 Census: An Introduction,” November 2021, U.S. Census Bureau, Washington DC https://www.census.gov/library/publications/2021/decennial/2020-census-disclosure-avoidance-handbook.html

U.S. Census Bureau (2022a) “Understanding Disclosure Avoidance- Related Variability in the 2020 Census Redistricting data, “ U.S. Census Bureau, Washington DC. January 28. https://www.census.gov/library/fact-sheets/2022/variability.html

U.S. Census Bureau (2022b) “Revised Data Metrics for 2020 Disclosure Avoidance,” U.S. Census Bureau, Washington DC.

U.S. Census Bureau (2022c) Post-Enumeration Survey and Demographic Analysis Help Evaluate 2020 Census Results, March 10 , Census Bureau Releases Estimates of Undercount and Overcount in the 2020 Census

U.S. Census Bureau (2022d) Detailed Summary Metrics

U.S. General Accountability Office (2020). “COVID-19 Presents Delays and Risks to Census Counts,” U.S. General Accountability Office, Washington, DC., https://www.gao.gov/products/GAO-20-551R .

Vink, J. (2019). “Elementary School Enrollment,” https://www.nationalacademies.org/event/12-11-2019/workshop-on-2020-census-data-products-data-needs-and-privacy-considerations .

White House Equitable Data Working Group (2022) “A Vision for Equitable Data : Recommendations from the Equitable Data Working Group,” https://www.whitehouse.gov/wp-content/uploads/2022/04/eo13985-vision-for-equitable-data.pdf

State of preschool 2021. https://nieer.org/wp-content/uploads/2021/08/YB2020_Full_Report_080521.pdf

4. Deborah Stein, Partnership for America's Children

PARTNERSHIP FOR AMERICA'S CHILDREN

State and local child advocates working
together for a better future

Comments on the implications of DP in the March 2022 Demonstration File

On behalf of the Partnership for America's children, I am submitting these comments on the implications of the March 2022 differential privacy demonstration product. I submit these comments after reviewing the research of Dr William O'Hare, which raises significant concerns.

The Partnership's mission is to support its network of state and community multi-issue child advocacy organizations in effective advocacy. The Partnership has 49 member organizations in 40 states that advocate to improve policies for children at the state, local and federal level. Collectively they represent over 90% of the nation's children. Partnership members use Census data in their advocacy, and thirty Partnership members are also KIDS COUNT grantees in their state, serving as that state's data hub on children for policy makers, administrators, and nonprofits.

The Partnership for America's Children served as the national hub on the undercount of young children in the 2020 Decennial Census. In this role the Partnership formed and continues to co-lead a national working group of child-serving organizations that is working to improve the count of young children in all Census Bureau demographic products.

Our concerns about the differential privacy demonstration product are two fold.

First, Dr O'Hare's research shows that at sub-state geographies, the error rate created by the use of differential privacy on data for young children are significant. This is particularly important when federal agencies use this data to allocate federal funds directly to localities, such as Local Education Authorities, especially for special education and early intervention services, and for the expansion of Head Start, because those funding formulas include data on the number of young children. We note however that Title 1 funds are based on the number of children age 5 to 17 and that children age 5 to 9 are the second most undercounted age group. While Dr O'Hare's research didn't look at the error rate for that age group we are concerned that it might also be significant, adding to the data problems for that age group, given the significant error rate for 4 year olds. The problems from DP also will have significant implications in states that use census data to allocate federal funds to localities after the state gets an allocation based on the state level data. We do not know how often that happens but are concerned about it particularly in the allocation of funds for child care and WIC.

Second, we are very concerned about the use of an algorithm that breaks the relationship between children and parents because so much of what we need to know about children is how their families are structured: Do they live with both parents, one parent, grandparents, foster parents? This data is completely unavailable from the 2020 census in the DHC file. We understand that the algorithm for the

DDHC files should not have this problem, but we are not certain that the Bureau has in fact developed a final algorithm that will protect that relationship. We want to emphasize that this is a significant concern, and that it would be even more of a problem should the ACS privacy algorithm have such a break. That is because the ACS is our most important source of child poverty data and child poverty is calculated based on household income, not the individual child's income. We need child poverty data from the ACS, and we need it to continue to be calculated in the same way so we have continuity of data.

Please feel free to contact me if you have any questions about our concerns

Sincerely,

Deborah Stein
Network Director

5. Rachel Cortes, Claritas

This email is in response to the request for feedback on the latest DHC demonstration files from the 2010 decennial census released in March 2022. After evaluating the fitness-for-use of the data, Claritas would like to provide comments on the potential issues with the data.

Claritas is a private data company and has relied for decades on decennial census data at the block-group level by demographic characteristics as a vital input to our yearly population estimates and 5-year projections. Our products are used as an input to numerous data products by companies across the country for countless uses by large and small businesses in every industry as well as non-profit organizations and local governments. Use cases for these products are wide-reaching and varied and errors in census data (and resulting error in Claritas estimates) impair the ability of businesses to make informed decisions, with potentially negative consequences for communities.

Having accurate and reliable data at the block group-level is vital to our products—census tract-level is not sufficient. Claritas products require data at the block-group level because small area data are needed to make community-level decisions by businesses. We urge the census bureau to make all tables that were available at the block group-level in 2010 also available at the block group-level in 2020.

In general, it appears that a problem with the Top Down Algorithm is that it does not synchronize data for population and data for households, and the resulting inconsistencies and errors appear to impact DHC data that are critical for Claritas products and numerous products which use them as an input. If things like more householders of a certain age than population of that age in a given block group occur in the data, we either have to perpetuate that error by leaving it or alter it and thereby possibly creating a different error in the underlying data.

Even if the population by age and householders by age data are reasonably accurate, the lack of synchronization between them is a problem, such as when a small area shows more householders in an age category than persons in that category. And even where the data do not violate that simple condition of consistency, the headship rates can still have large errors. Other DHC data that are critical to Claritas include households by type and number of persons and relationship to householder which also seem to be impacted by the TDA. Claritas urges the Census Bureau to continue to make improvements and concentrate efforts to make the household and person-level files consistent.

Another consequence of the implementation of the TDA causing impossible values in census tables is that it causes clients to doubt the value of census data and question our use of it in their products. An important part of our work is keeping clients informed on the data and methods that we use in our products. Values that even casual data users can identify as being incorrect lead to a lack of confidence in census bureau data and in our data by association. Improving the 2020 data products can help restore this confidence in census data.

Claritas appreciates the opportunity to comment on the 2010 demonstration data and we hope that the census bureau will continue to refine the application of the TDA so that when it is used for the 2020 DHC and detailed DHC products the number and magnitude of improbable and impossible values like those shown in these demonstration files will be reduced and overall accuracy and reliability will be increased making these data a dependable source for information about the population of the United States.

Thank you,

Rachel

Rachel Cortes
Senior Demographer

6. Jan Vink, Ph.D., Cornell Program on Applied Demographics

Feedback on the demonstration data sets released in March/April 2022

Authors: Jan Vink and Leslie Reynolds

Cornell Program on Applied Demographics
May 16, 2022

Email: padinfo@cornell.edu

1 INTRODUCTION AND CONCLUSIONS

On March 16’th the Census Bureau released a new set of demonstration data with person level characteristics, followed by household (unit) characteristics on April 14’th.

General impressions and conclusions:

- The last demonstration data set that contained detailed age information was in May 2020, and the first demonstration data set in October 2019 was the last one that contained information on households. Many changes to the algorithms were made since that time and the PLB was increased significantly. Therefore, we make no attempt to compare this release with the earlier releases.

- The Census Bureau already recognized some shortcomings in the sex/age distributions, but besides those, the age/sex distributions for the total population look usable for on-spine geographies (blocks groups and above) and geographies that build off optimized block groups. Differences in median age get smaller with growing population size.
- Differences in median age for off-spine geographies (like CDPs and School Districts) are much bigger than for the on-spine geographies. While CDPs might not have their own general government, county governments have delineated them to provide better governance for these clusters of population and need more accurate age pyramids.
- Breaking the link between persons and households generates many inconsistencies in the data. Census data users use various indicators that rely on the data of both households and persons:
  - Persons per Household (household population divided by number of households),
  - headship rates (householders of a certain age divided by household population of that age) are used to link age structures of the population with number of households,
  - minority home ownership (householders of a minority group divided by population of that group)

  Inconsistencies in the data require data users to derive alternative estimates that are feasible, but not longer solely based on area specific Census Data. These inconsistencies can also be seen as symptoms of distortions of the underlying distributions.

  Inconsistencies happen on all levels of geography analyzed (county and below)

- Table cells for household type and household size that generally have lower counts (less common household types and household sizes) often have very large percentage errors (e.g. over 30% of tracts have more than 10% error) which severely limits the usability of these tables.
- Various selection biases that are not well-understood exist within these data. For example: areas with mostly rental occupied houses have a positive bias for households with children, whereas areas with mostly owner occupied houses have a negative bias for households with children.

Analyses are based on the data downloaded directly from the Census Bureau servers or on data downloaded from NHGIS-IPUMS[1]. Several of the analyses are limited to New York State.

[1] David Van Riper, Tracy Kugler, and Jonathan Schroeder. IPUMS NHGIS Privacy-Protected 2010 Census Demonstration Data, version 20220310 [Database]. Minneapolis, MN: IPUMS. 2020.

2 ON- AND OFF-SPINE DIFFERENCES IN MEDIAN AGE

2.1 RESEARCH QUESTION:

How do differences in median age vary by type of geography, size of population, sex and race/ethnicity?

2.2 CONCLUSIONS:

- Very small population groups (< 200) seem to have a significant negative bias
- Unincorporated places (CDPs) have larger differences compared to Incorporated places (villages and cities). Counties create CDPs with the purpose of being able to govern over those communities like their incorporated counterparts.
- CDPs have larger differences than School Districts. Maybe because CDPs are more urban?
- CDPs below 2,000 population often have absolute difference of 2 year or more. This is rare among other geographies of size 500 or more
- The difference between CDPs and villages is most prominent under NH White Alone populations, not as much difference for Black Alone or Hispanic populations
- For population groups larger than 1,000 the mean absolute difference is mostly less than 1 year
- The mean absolute difference in median age for males or females where there are 1,000 to 2,000 males or females is smaller then total populations of 1,000 to 2,000
- The mean absolute difference for groups of two or more races that are 1,000 to 2,000 in size are larger than for other race categories that size

2.3 METHOD:

I used NHGIS-IPUMS data on places, sub-counties and unified school districts and used the name to determine type of geography. I only looked at geographies in New York State.

Types of geographies considered are:

- Cities (they are places and minor civil divisions in New York),
- Villages (incorporated),
- CDPs,
- Towns (MCDs),
- American Indian Reservations and
- Unified School Districts (SD).

Race groups coding:

1. Total population
2. White Alone (A)
3. Black Alone (B)
4. American Indian and Alaskan Native Alone (C)
5. Asian Alone (D)
6. Native Hawaiian and Pacific Islander Alone (E)
7. Other Race Alone (F)
8. Two or more races (G)
9. Hispanic (H)
10. Non-Hispanic White Alone (I)

Sex coding:

1. All
2. Male
3. Female

For each of the subgroup I looked at the population size in SF1 and coded that in the following population size bins:

- 0-9 (Excluded from further analyses)
- 10-199
- 200-499
- 500-999
- 1,000-1,999
- 2,000-4,999
- 5,000-9,999
- 10,000+

2.4 Metrics

For each subgroup I calculated the difference in median age as the median age in the demonstration data minus the median age in the SF1.

For each subgroup based on geography, size of population, sex and race/ethnicity I calculated a Mean Difference (possibly indicating bias) and Mean Absolute Difference (measure of accuracy). I also tallied the number of times the absolute difference exceeded 1, 2 and 5 years which allows me to created percentage of all cases that exceeds those thresholds.

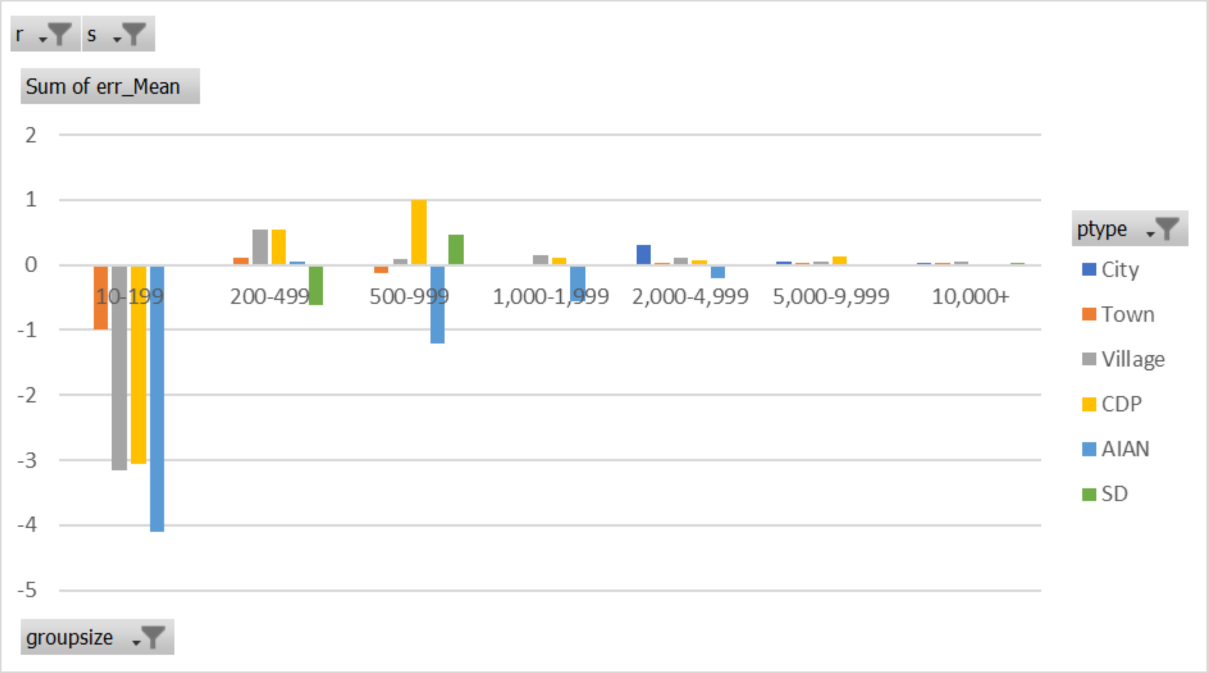
2.5 Results

2.5.1 Total population by geography type and size

Number of geographies by type and size

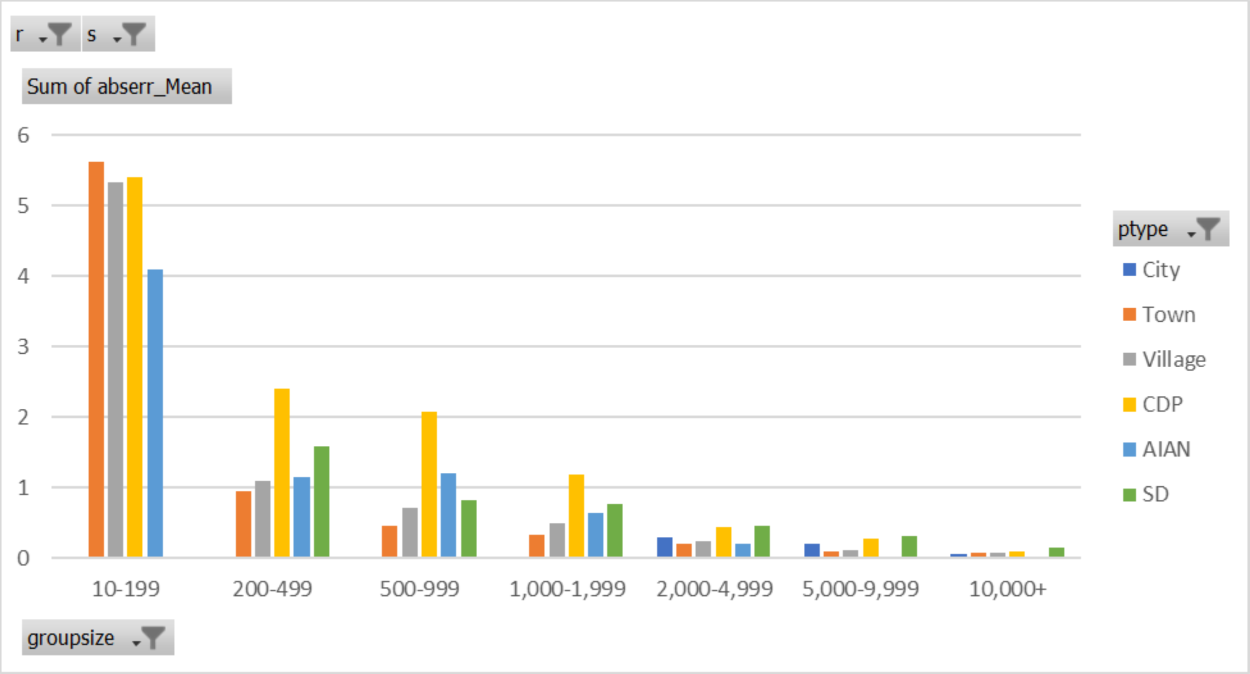
groupsize	City	Town	Village	CDP	AIAN	SD	Grand Total
10-199		6	6	20	2	1	35
200-499		25	67	67	4	8	171
500-999		75	111	90	1	5	282
1,000-1,999		223	125	97	3	18	466
2,000-4,999	2	296	137	121	1	118	675
5,000-9,999	8	151	74	73		198	504
10,000+	52	156	35	104		319	666
All	**62**	**932**	**555**	**572**	**11**	**667**	**2799**

Figure 1: Mean difference in median age



Conclusion: Very small population groups (< 200) seem to have a significant negative bias

Figure 2: Mean absolute difference in median age



Conclusion: Unincorporated places (CDPs) have larger differences compared to Incorporated places (villages and cities).

Conclusion: CDPs have larger differences than School Districts. Maybe because CDPs are more urban?

Figure 3: Fraction of geographies with an absolute difference of 1 yr or more

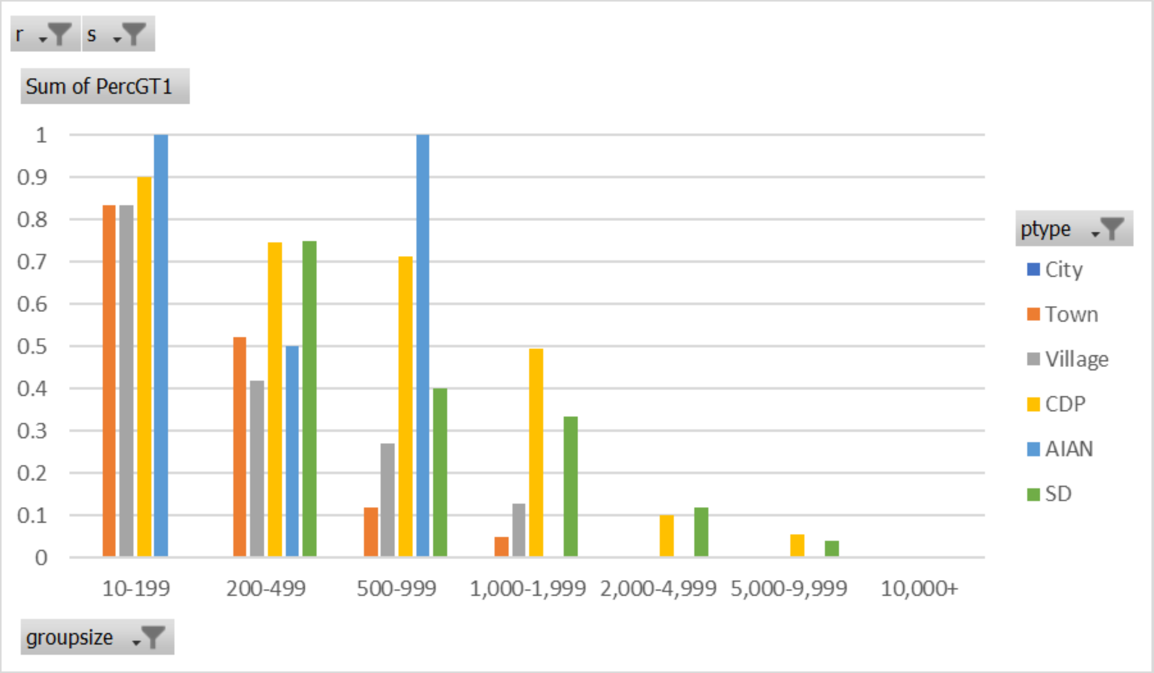
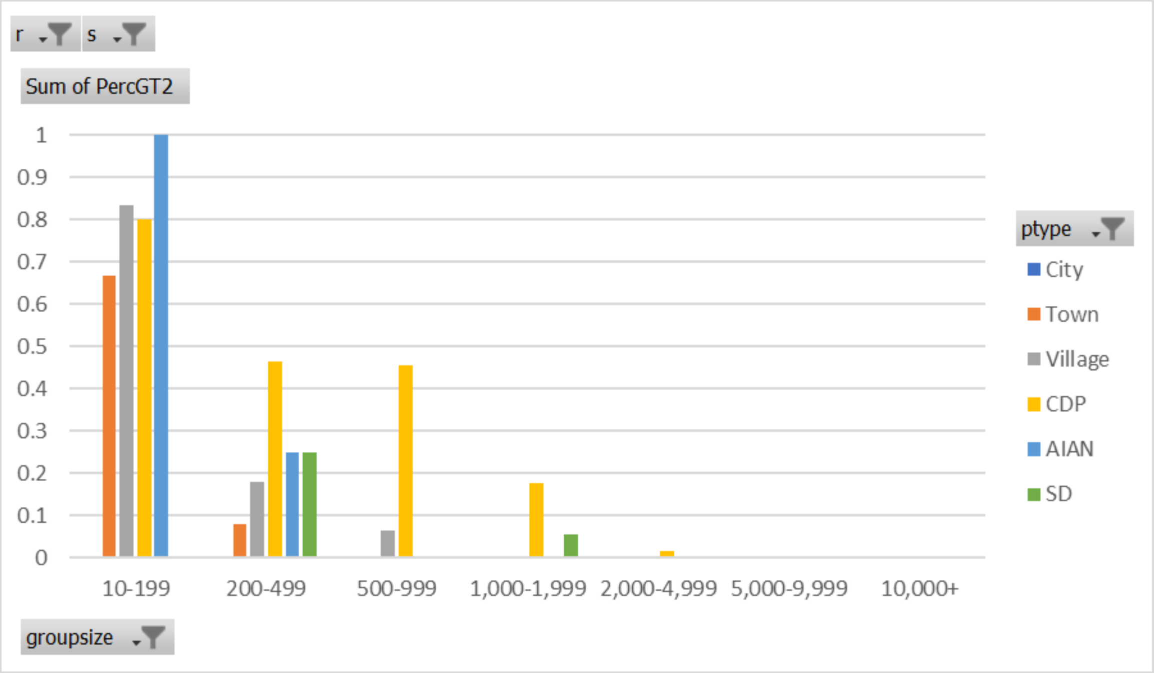
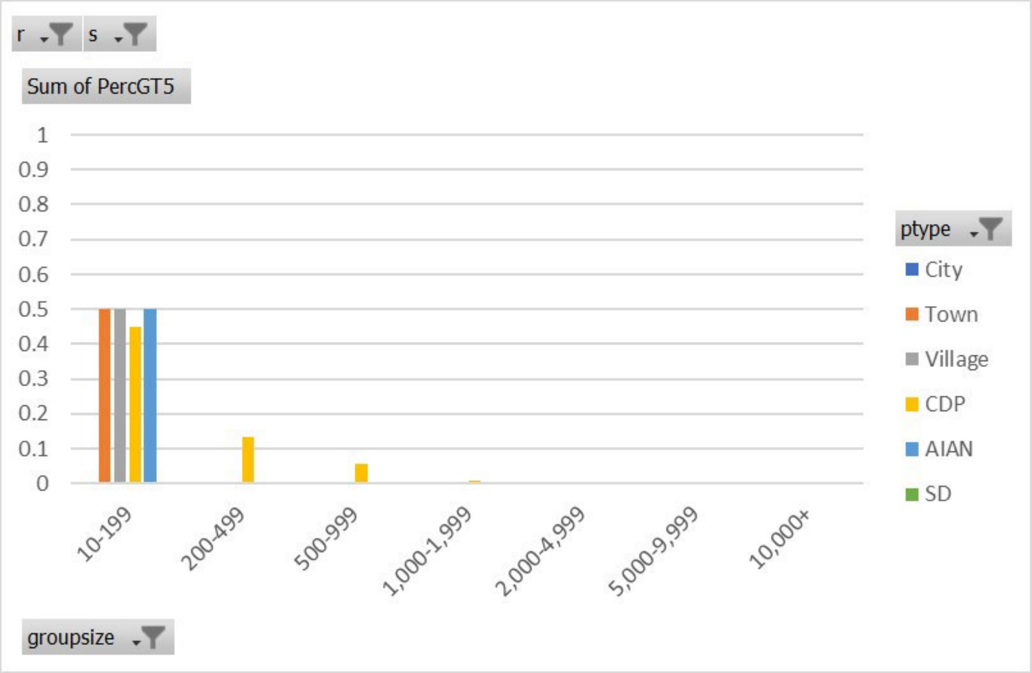


Figure 4: Fraction of geographies with an absolute difference of 2 yr or more



Conclusion: CDPs below 2,000 population often have absolute difference of 2 year or more. This is rare among other geographies of size 500 or more

Figure 5: Fraction of geographies with an absolute difference of 5 yr or more



2.5.2 Mean absolute differences for major race/ethnicity groups

*Figure 6: Mean absolute difference for **Non-Hispanic White Alone** populations*

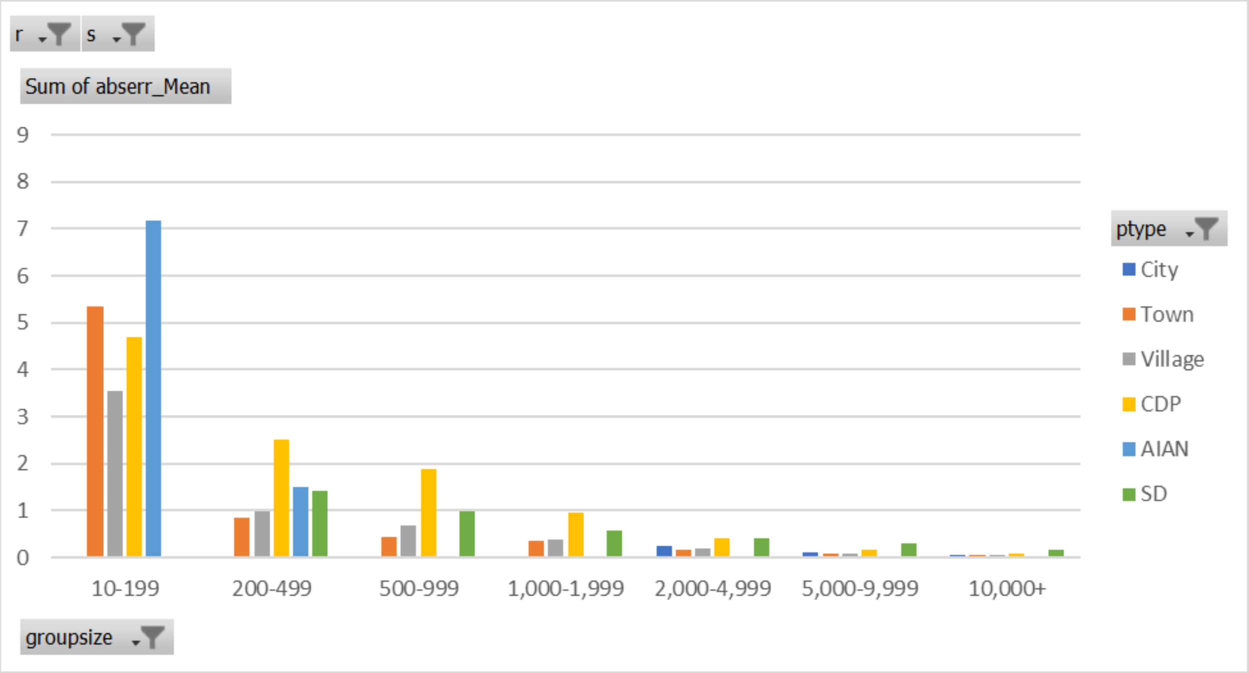


*Figure 7: Mean absolute difference for **Black Alone** populations*

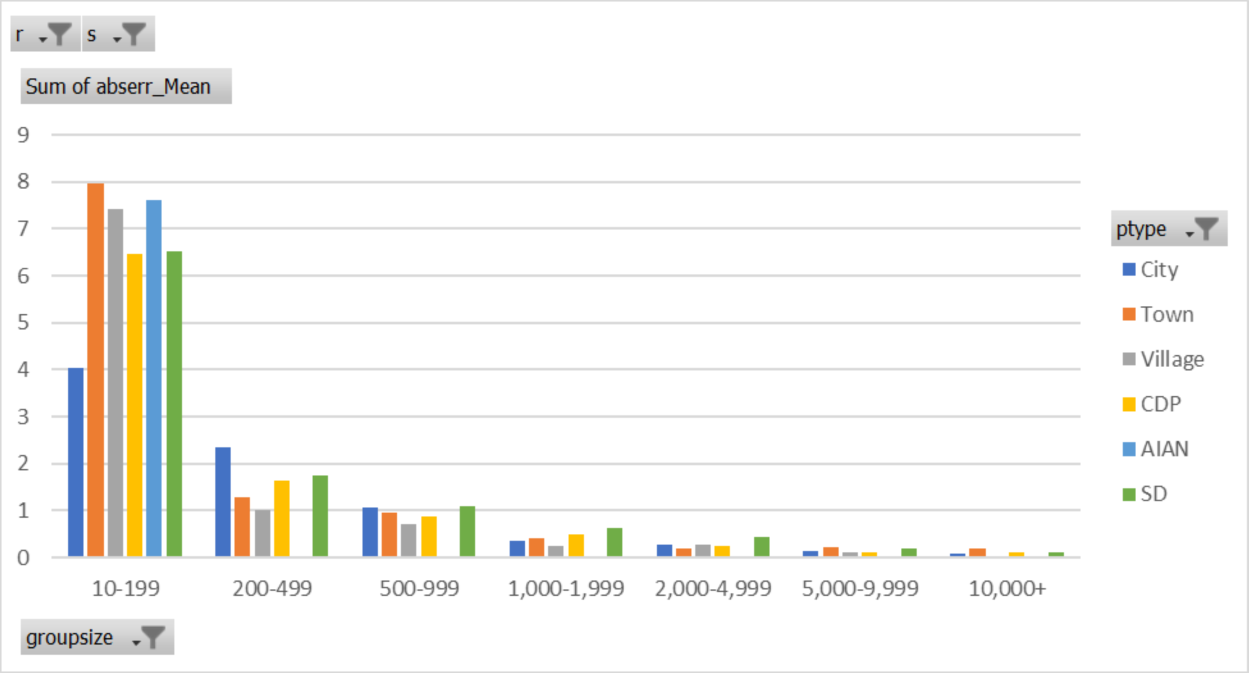
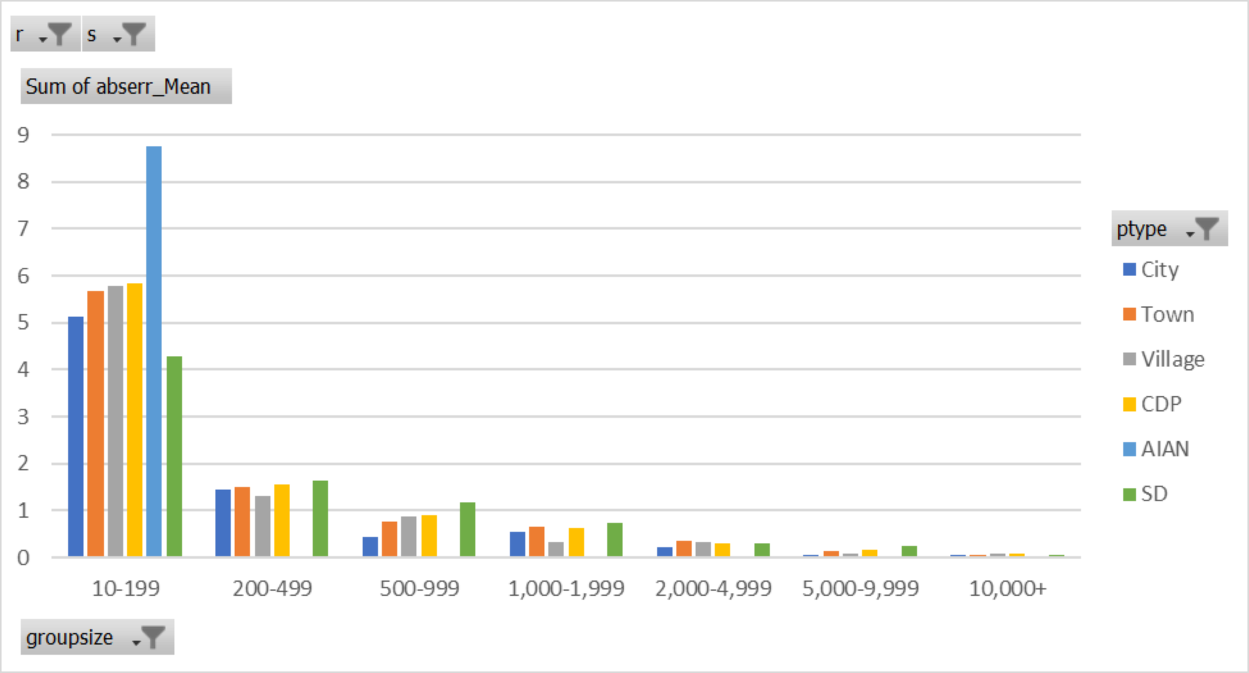


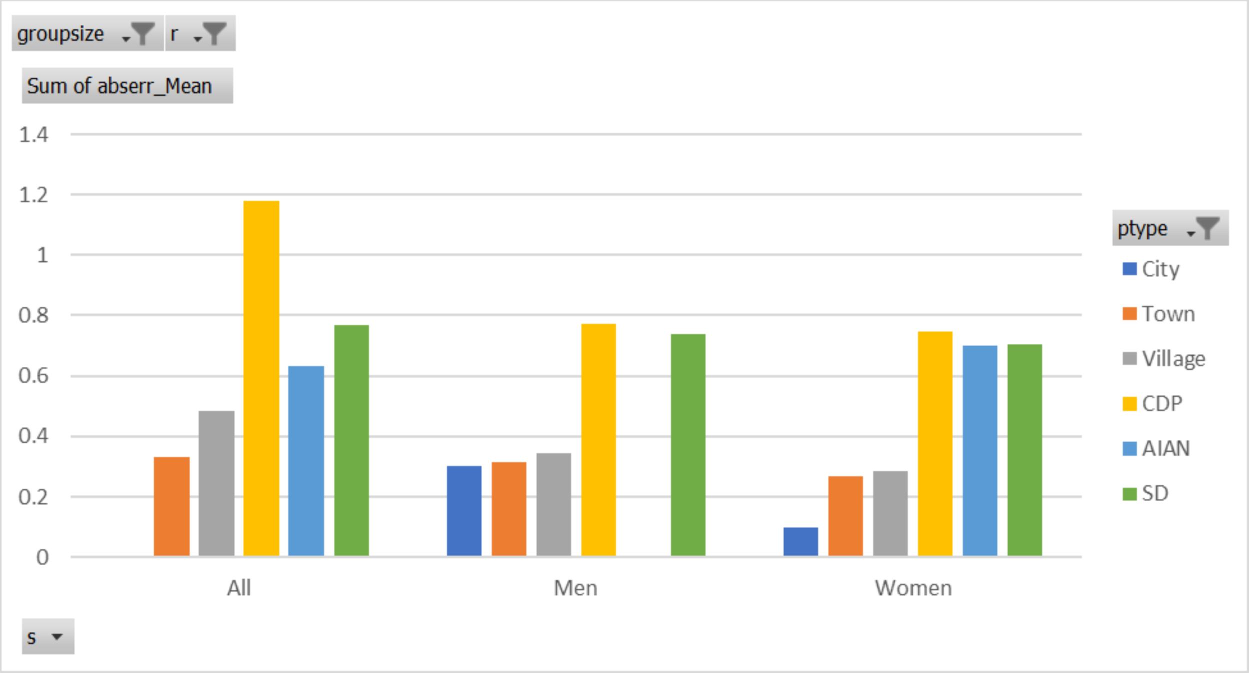
*Figure 8: Mean absolute difference for **Hispanic** populations*



Conclusion: The difference between CDPs and villages is most prominent under NH White Alone populations, not as much difference for Black Alone or Hispanic populations

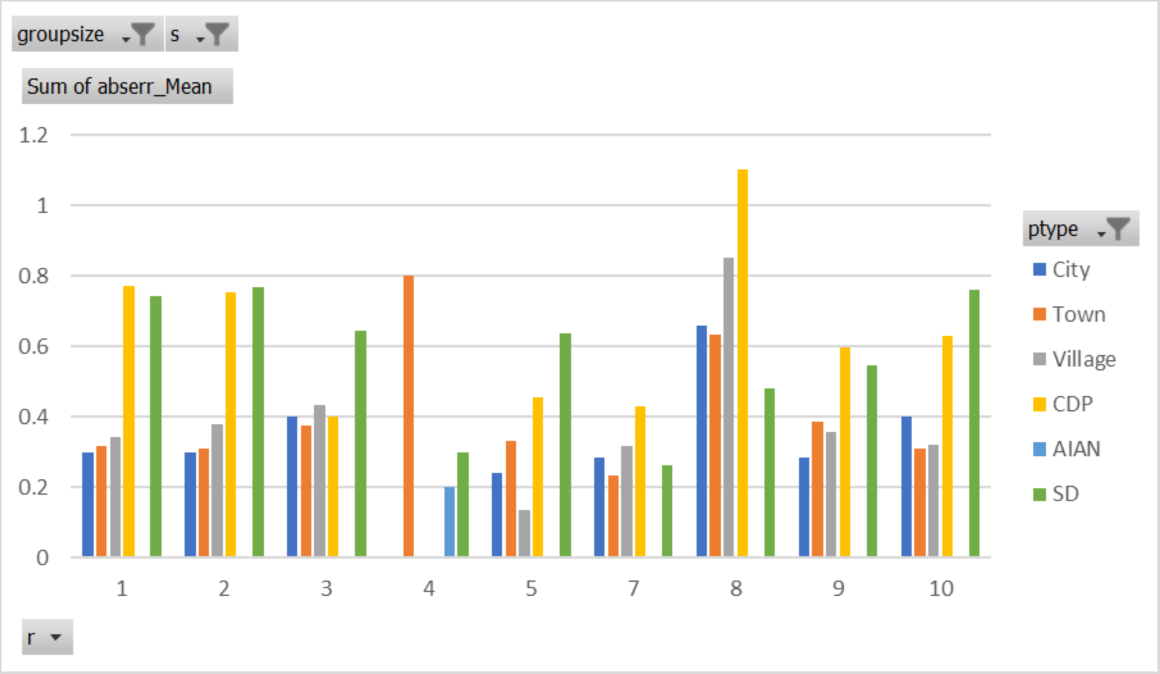
2.5.3 Mean absolute differences for population sizes 1,000 – 1,999

Figure 9: Mean absolute differences by sex



Conclusion: The mean absolute difference in median age for males or females where there are 1,000 to 2,000 males or females is smaller then total populations of 1,000 to 2,000

Figure 10: Mean absolute differences by race



Conclusion: The mean absolute difference for groups of two or more races (race code 8) that are 1,000 to 2,000 in size are larger than for other race categories that size

3 GEOGRAPHIES WITH IMPOSSIBLE STATISTICS

3.1 RESEARCH QUESTION:

Breaking the connection between households and persons lead to many impossible statistics. How often do they appear for different levels of geography?

3.2 CONCLUSIONS:

- Block level data is full of inconsistencies
- There are a few fields that by definition should be the same in the person file as in the household file (e.g. householders = households, householders living alone = single person households). In this data set, this is very rarely the case. Big differences also occur rather often in sub-county geographies
- The number of older age householders often exceeds the population in that age group
- The number of minority householders often exceeds the population in those race groups. At the block group level the householders often outnumber the population by more than 10 and more than 10%

3.3 METHOD:

The person file and household file were merged. The SUMLEV field was used to determine the geographic level. Records with zero population were left out of the analyses. Several inconsistencies were flagged and if the difference exceeded 10 and 10% were flagged as a big error. Extreme examples were chosen by looking at a combination of the percent error and the size of the population

3.4 MORE HOUSEHOLDS THAN HOUSEHOLD POPULATION

The household population (from table H8) should be larger than the number of occupied houses (from table H3).

Summary level	N	Flagged count	%	Big error count	%
County	62	0		0	
Tract	4870	1	0.02%	0	
Block group	15194	2	0.01%	0	
Blocks	244281	4646	1.9%	84	0.03%
MCD	1010	0		0	
Place	1189	1	0.08%	0	
Unified SD	669	0		0	

Extreme examples:

Block 36109613001000: Total population = 299, Household population = 1, occupied houses = 15

Block 360550094002030: Total population = 140, Household population = 140, occupied houses = 156

3.5 HOUSEHOLD POPULATION WITHOUT OCCUPIED HOUSES

If there is household population (from table H8) than the number of occupied houses (from table H3) should be non-zero.

Summary level	N	Flagged count	%
County	62	0	
Tract	4870	10	0.2%
Block group	15194	12	0.08%
Blocks	244281	16930	6.9%
MCD	1010	2	
Place	1189	0	
Unified SD	669	0	

Extreme examples:

Block 361031456033001:	Total population = 71, Household population = 71, occupied houses = 0 (out of 15 total housing units)

3.6 HOUSEHOLDERS NOT EQUAL TO HOUSEHOLDS

The population with relationship “householder” (from table P19) should be equal to the number of occupied houses (from table H3).

This is especially important when calculating Persons per Household where we often have two different numbers for the denominator.

Summary level	N	Flagged count	%	Big error count	%
County	62	62	100%	0	
Tract	4870	4806	98.7%	112	2.3%
Block group	15194	14957	98.4%	2763	18.2%
Blocks	244281	222704	91.2%	44501	18.2%
MCD	1010	996	98.6%	23	2.3%
Place	1189	1173	98.7%	188	15.8%
Unified SD	669	665	99.4%	9	1.3%

Extreme examples:

New Cassel CDP:	Total population = 14,056, Householders = 3,316, occupied houses = 2,973
Quogue village:	Total population = 1,004, Householders = 374, occupied houses = 458
Blockgroup 360470776003:	Total population = 1,119, Householders = 463, occupied houses = 319

3.7 HOUSEHOLDERS LIVING ALONE FROM THE PERSON FILE NOT EQUAL TO HOUSEHOLDERS LIVING ALONE FROM THE UNIT FILE

The population with relationship “householder living alone” (from table P19) should be equal to the number of households with household type “Householder living alone” (from table P16).

Summary level	N	Flagged count	%	Big error count	%
County	62	62	100%	1	1.6%
Tract	4870	4756	97.7%	1060	21.8%
Block group	15194	14884	98.0%	8445	55.6%
Blocks	192337	222704	78.7%	24182	9.9%
MCD	1010	986	97.6%	118	11.7%
Place	1189	1154	97.0%	309	26.0%
Unified SD	669	660	98.7%	46	6.9%

Extreme examples:

Brentwood UFSD: Total population = 87,297, Householders living alone 1355 male + 1564 female = 2,919, Household type “householder living alone” = 2,449

Blockgroup 360550116013: Total population = 1362, Householders living alone 64 male + 121 female = 185, Household type “householder living alone” = 57

Block 360610075001004: Total population = 413, Householders living alone 7 male + 31 female = 38, Household type “householder living alone” = 166

3.8 HOUSEHOLD POPULATION UNDER 18 LESS THEN NUMBER OF HOUSEHOLDS WITH CHILDREN UNDER 18

The household population under age 18 (from table P15) should be at least as large as the number of households with one or more people under 18 (from table P21).

Summary level	N	Flagged count	%	Big error count	%
County	62	0		0	
Tract	4870	29	0.6%	6	0.1%
Block group	15194	124	0.8%	48	0.3%
Blocks	192337	42031	17.2%	1032	0.4%
MCD	1010	2	0.2%	2	0.2%
Place	1189	12	1.0%	0	
Unified SD	669	2	0.3%	1	0.015%

Extreme examples:

Block group 360670163004: Total population = 750, Household population under 18 = 75, Households with children under 18 = 101

3.9 NOT ENOUGH HOUSEHOLD POPULATION TO FILL THE HOUSEHOLD BY SIZE STATISTICS

One can calculate an under bound for the household population from table H9 (households by size) by multiplying each size category by the size and multiply the 7-or-more category by 7. The household population (table H8) should be larger than this under bound.

One can also calculate what the average household size of the 7 or more category should be to match the household population. Values much larger than 10 are very improbable. These analyses are not part of this feedback.

Summary level	N	Flagged count	%	Big error count	%
County	62	13	21.0%	0	
Tract	4870	1852	38.0%	30	0.6%
Block group	15194	6665	43.9%	399	2.6%
Blocks	192337	104879	42.9%	42484	17.4%
MCD	1010	470	46.5%	10	1.0%
Place	1189	617	51.9%	76	6.4%
Unified SD	669	277	41.4%	6	0.9%

Extreme examples:

Block 361059515002036: Household population = 4, 1-person households = 2, 7+ households = 9,

household population under bound based on household size = 65

Aurora village: Household population = 336, under bound based on housing size = 466 (55*1 + 59*2+25*3+17*4+16*5+7*6+4*7)

3.10 MORE HOUSEHOLDERS OF A CERTAIN AGE GROUP THAN POPULATION OF THAT AGE GROUP

The number of people in an age group (from table P12) should greater or be equal to the number of householders in that age group (from table H13)

Geographies without population and without householders in a certain age group are excluded from these analyses.

Figure 11: Share of geographies with number of householders exceeding population by age group

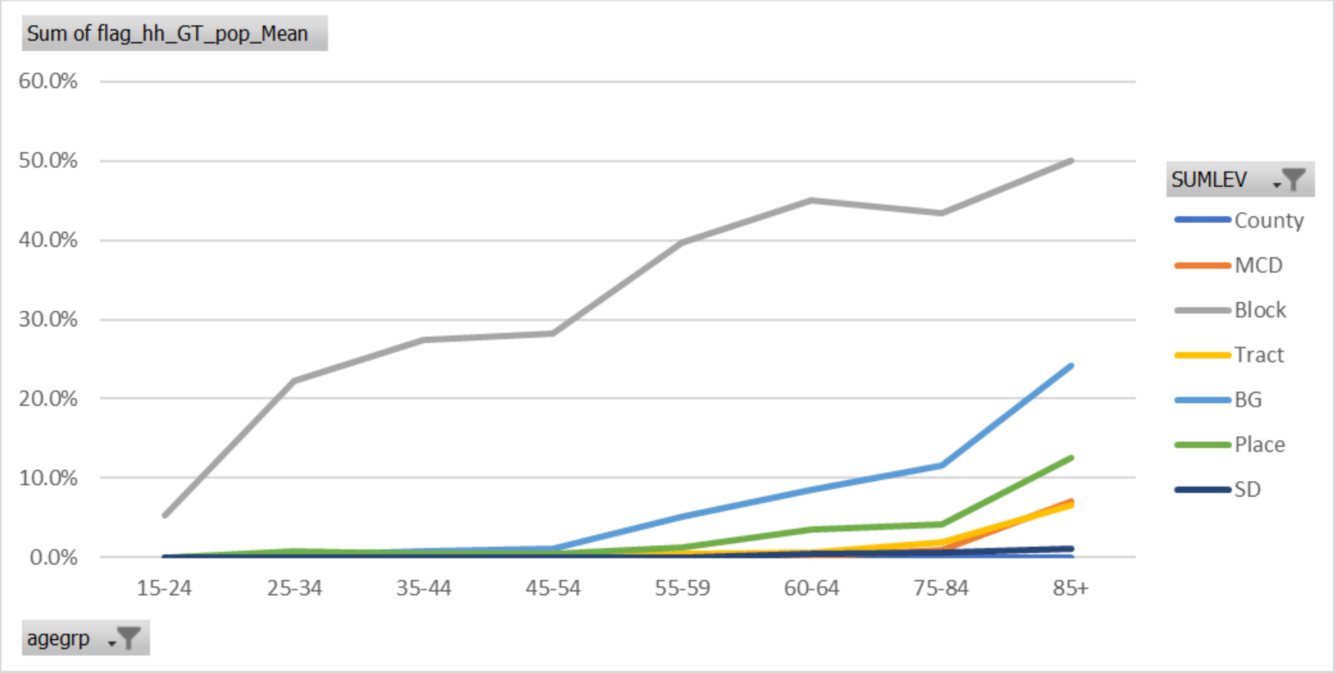
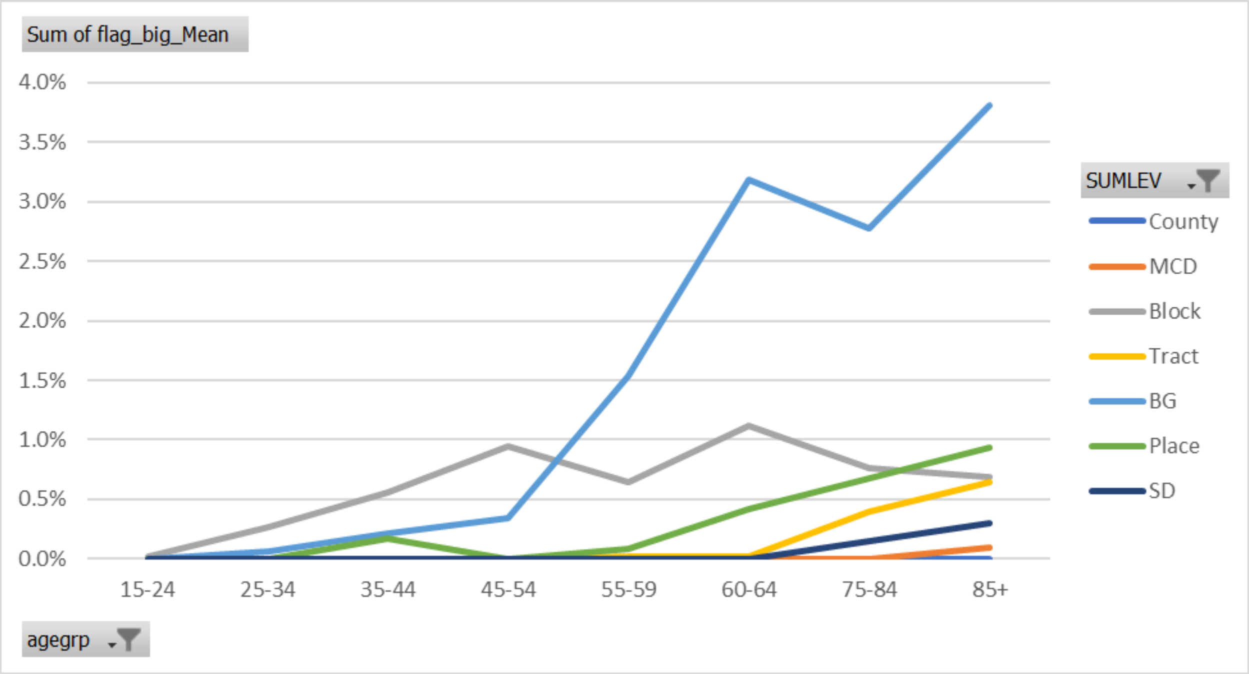

Sum of flag_hh_GT_pop_Mean
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
15-24
25-34
35-44
45-54
55-59
60-64
75-84
85+
agegrp
SUMLEV
County
MCD
Block
Tract
BG
Place
SD

Figure 12: Share of geographies with number of householders greatly exceeding population by age group



Extreme examples:

Blockgroup 360470890004: householders age 60-64 = 76, population age 60-64 = 7

3.11 More householders of a certain race/ethnicity group than population of that same group

The number of people in an age group (from table P12) should greater or be equal to the number of householders in that age group (from table H13)

Geographies without population and without householders in a certain race/ethnicity group are excluded from these analyses.

Figure 13: Share of geographies with number of householders exceeding population by race/ethnicity group

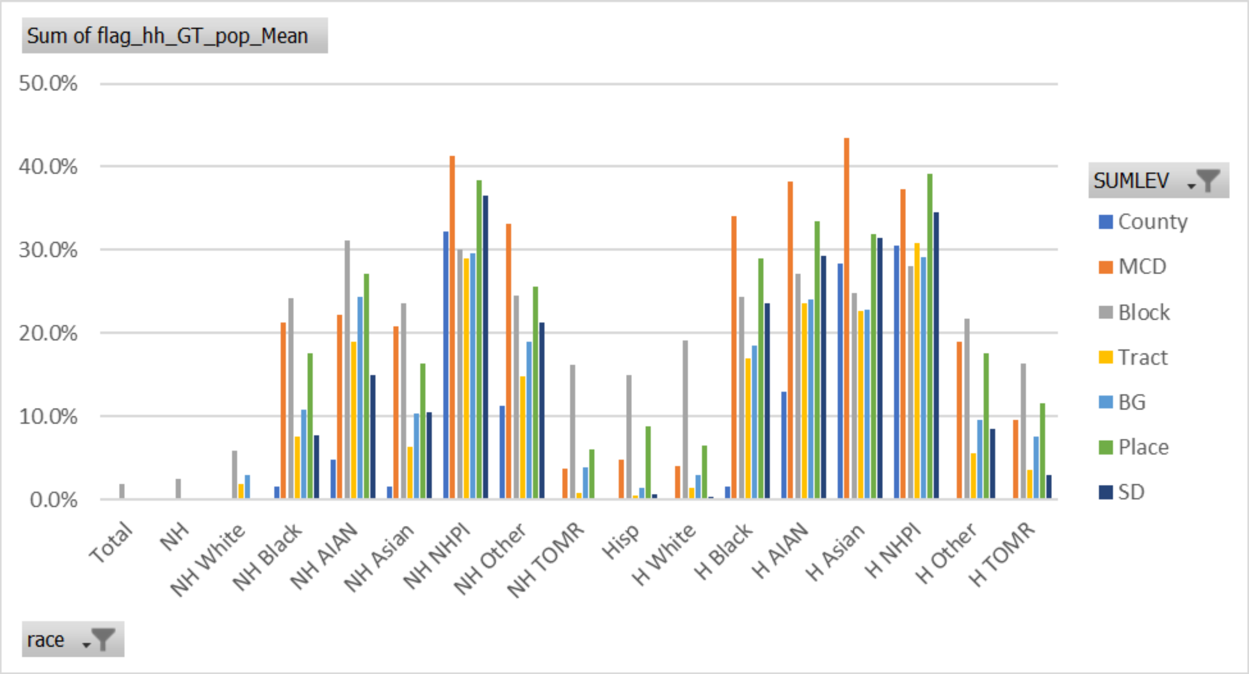
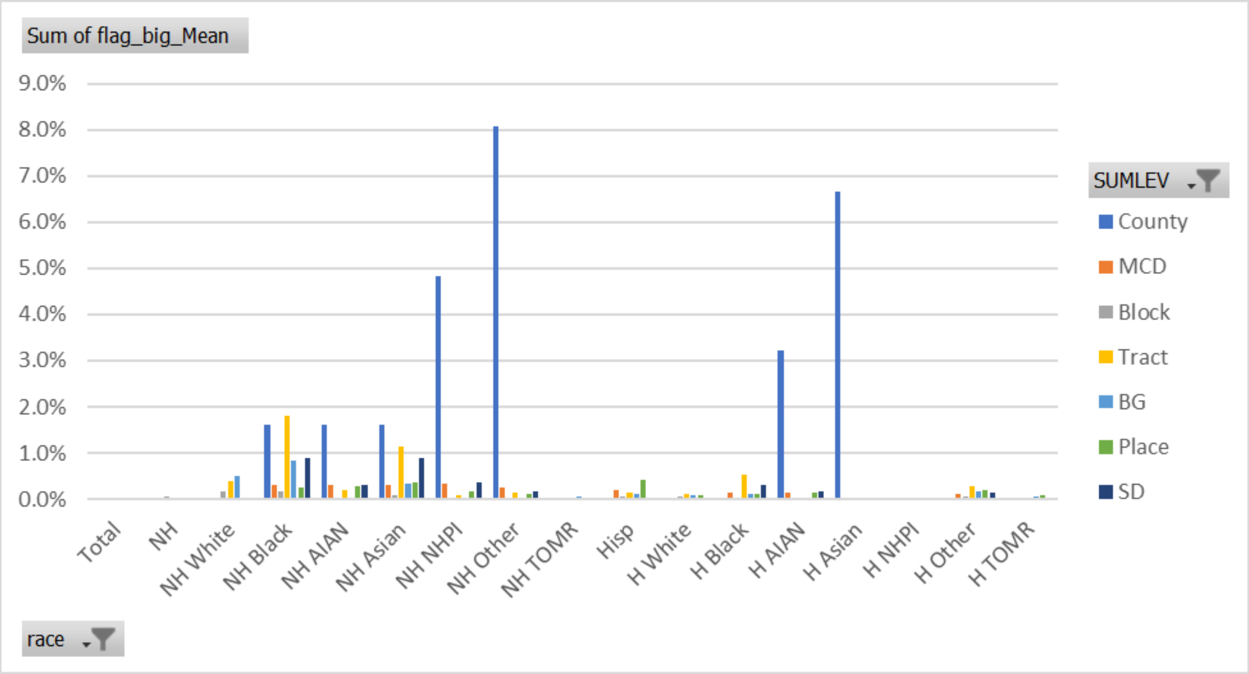


Figure 14: Share of geographies with number of householders greatly exceeding population by race/ethnicity group



Extreme examples:

Tract 36033940000: NH Black Alone householders = 33, NH Black Alone population = 1

4 COMPARING VILLAGES AND CENSUS DESIGNATED PLACES

Places are not on the traditional spine of the Top-Down Algorithm. Our understanding is that incorporated places are brought closer to the spine by creating an optimized blockgroup geography.

Furthermore, we understand that unincorporated places do not benefit the same optimization.

In this chapter results from incorporated places are compared with unincorporated places.

4.1 RESEARCH QUESTION:

Is there much difference in error metrics between incorporated places and unincorporated places

4.2 CONCLUSIONS:

- Most variables have more errors for CDPs than for villages of similar size 1,000 – 5,000
- Differences between CDPs and villages were most noticeable in the P12 tables on age and sex
- Tables on household type and household size showed big differences between SF1 and the demonstration data. There are also some differences between villages and CDPs

4.3 METHOD:

The NHGIS-IPUMS files do not contain information about their functional status. Instead I looked at the name of the geography; if it had a substring " village" I assumed this is an incorporated place and if it had a substring " CDP" I assumed an incorporated place. I further limited my analyses to places between 1,000 and 5,000 population in SF1 and at least 80% household population.

This resulted in 2897 CDPs (average 2371 persons in SF1) and 1005 villages (average 2173 persons in SF1)

I compared all variables in the person and housing unit file in the IPUMS data and chose three tables for further analyses.

4.3.1 Table P12: Population by age and sex

Figure 15: Average percentage error by age group (MALPE), by sex and village/CDP

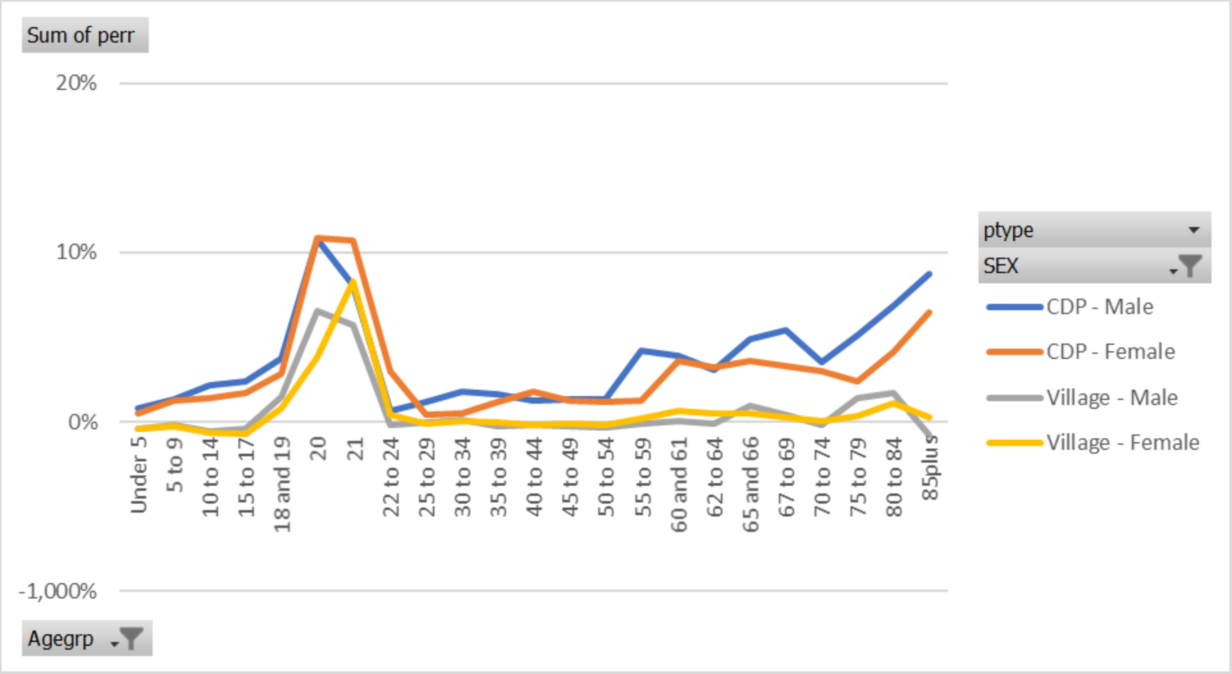
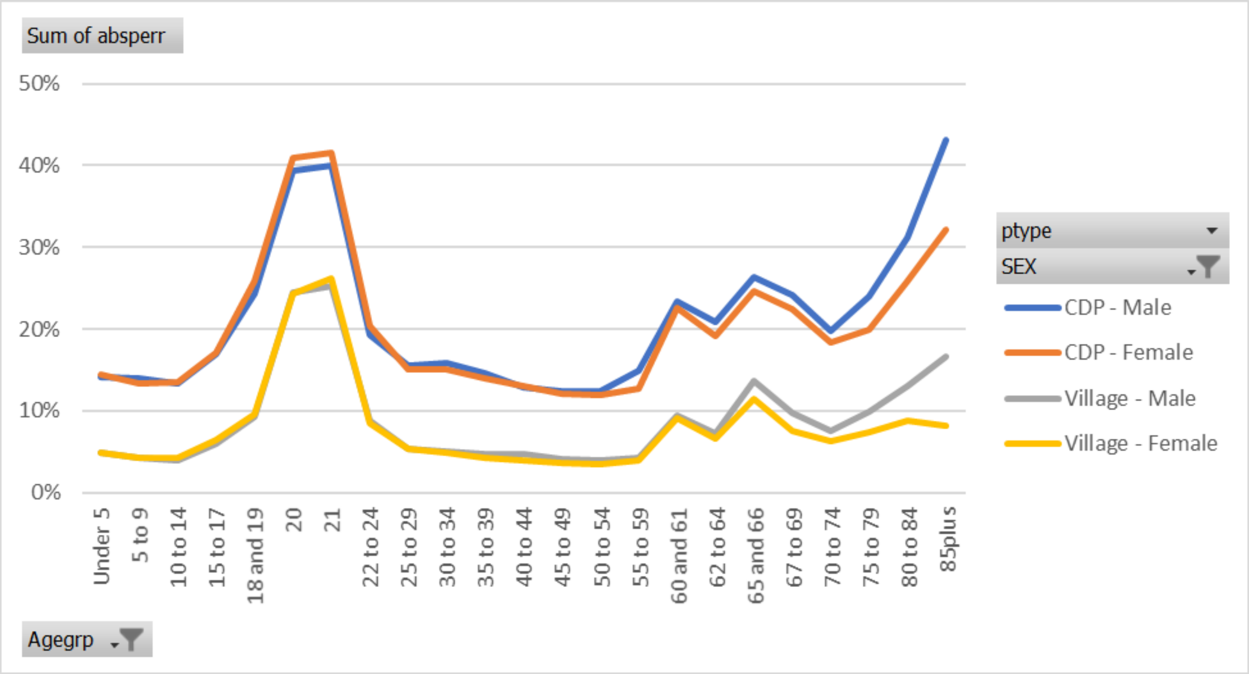

Sum of perr
20%
10%
0%
-1,000%
Under 5
5 to 9
10 to 14
15 to 17
18 and 19
20
21
22 to 24
25 to 29
30 to 34
35 to 39
40 to 44
45 to 49
50 to 54
55 to 59
60 and 61
62 to 64
65 and 66
67 to 69
70 to 74
75 to 79
80 to 84
85plus
Agegrp
ptype
SEX
CDP - Male
CDP - Female
Village - Male
Village - Female

Figure 16: Average absolute percentage error by age group (MAPE), by sex and village/CDP



4.3.2 Table P16: Household type

Codebook:

H8C001:	Total
H8C002:	Family households
H8C003:	Family households: Husband-wife family
H8C004:	Family households: Other family
H8C005:	Family households: Other family: Male householder, no wife present
H8C006:	Family households: Other family: Female householder, no husband present
H8C007:	Nonfamily households
H8C008:	Nonfamily households: Householder living alone
H8C009:	Nonfamily households: Householder not living alone

Figure 17: Average percentage error by household type (MALPE), by village/CDP

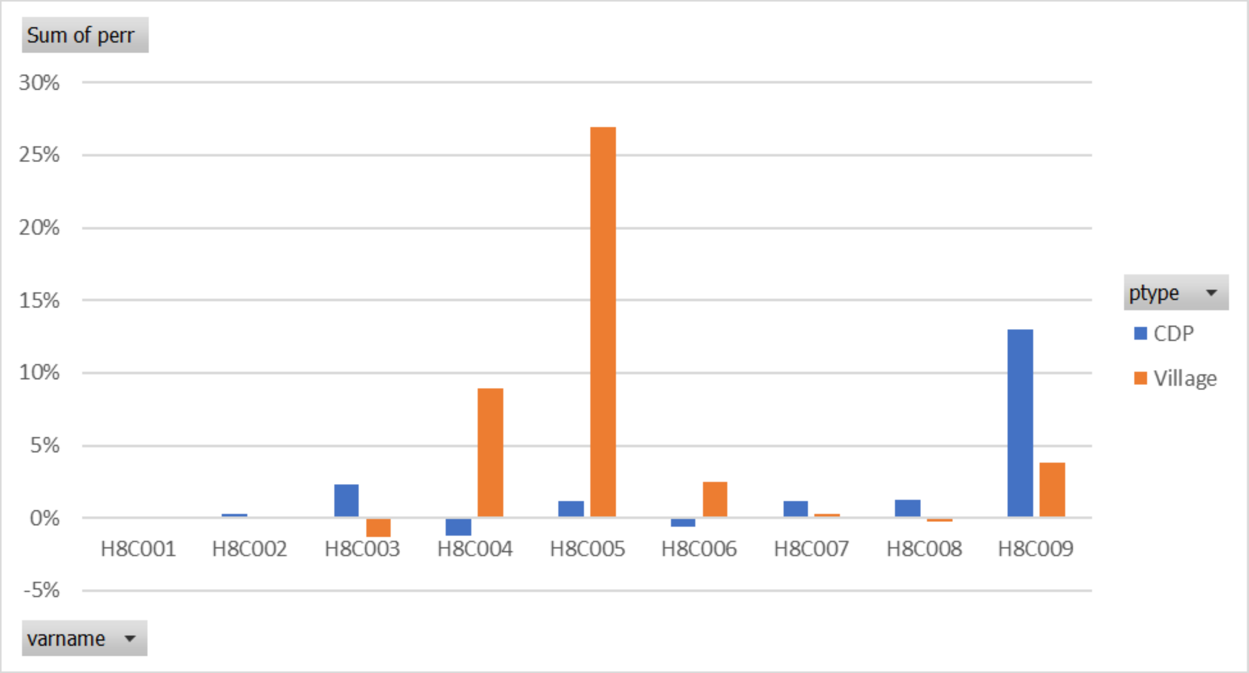
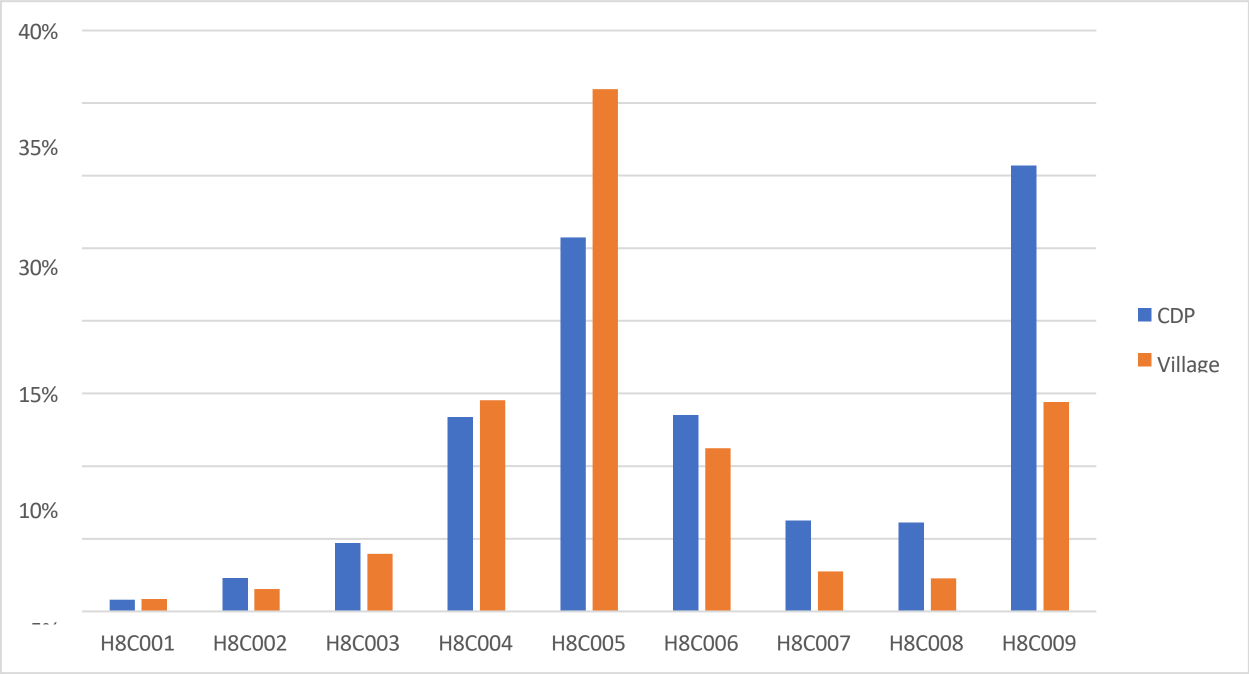


Figure 18: Average absolute percentage error household type (MAPE), by village/CDP



4.3.3 Table H13: Household size

Figure 19: Average error by household size (ME), by village/CDP

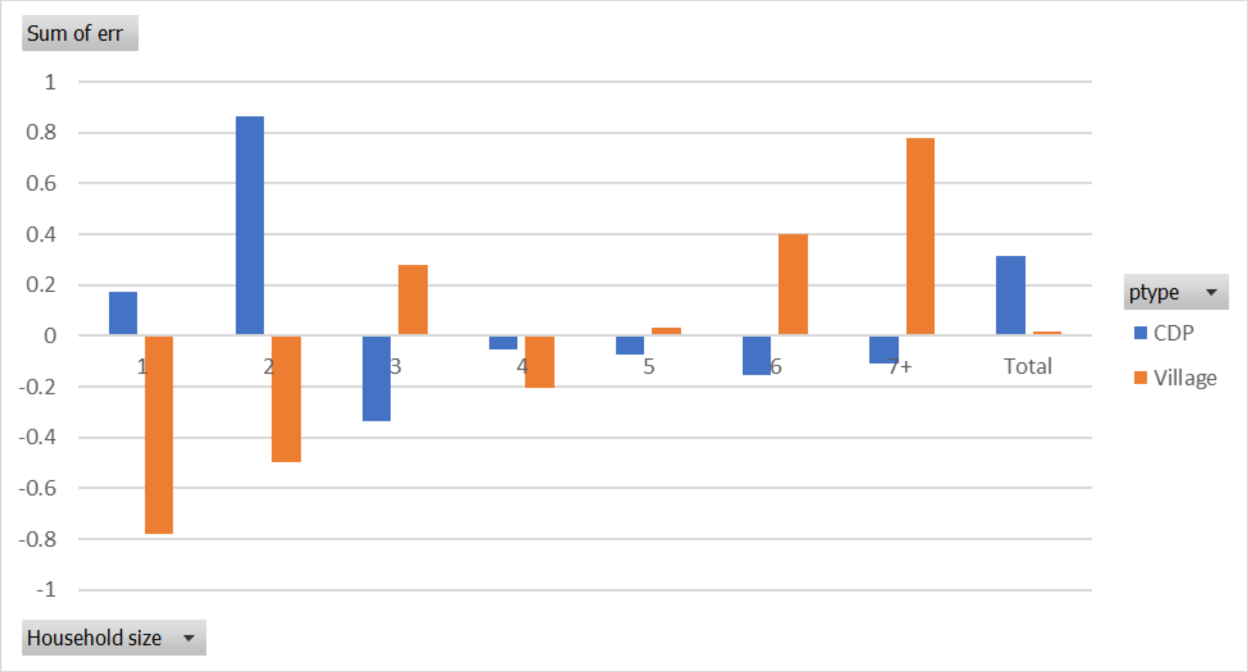


Figure 20: Average percentage error by household size (MALPE), by village/CDP

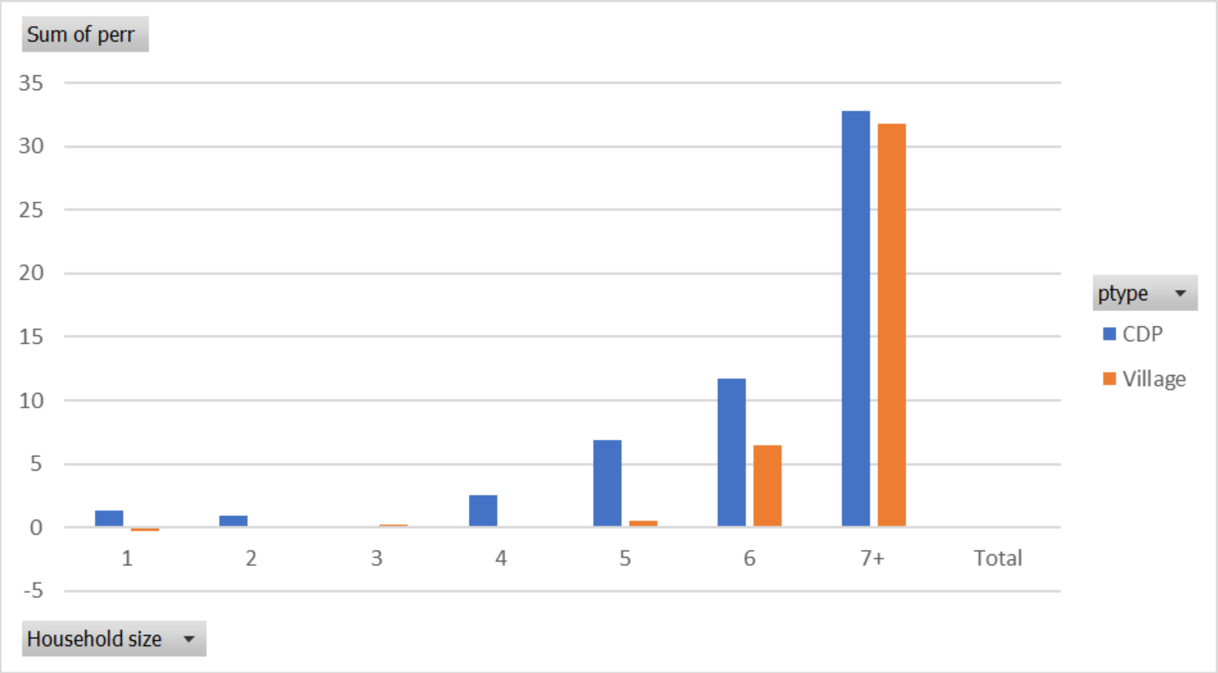
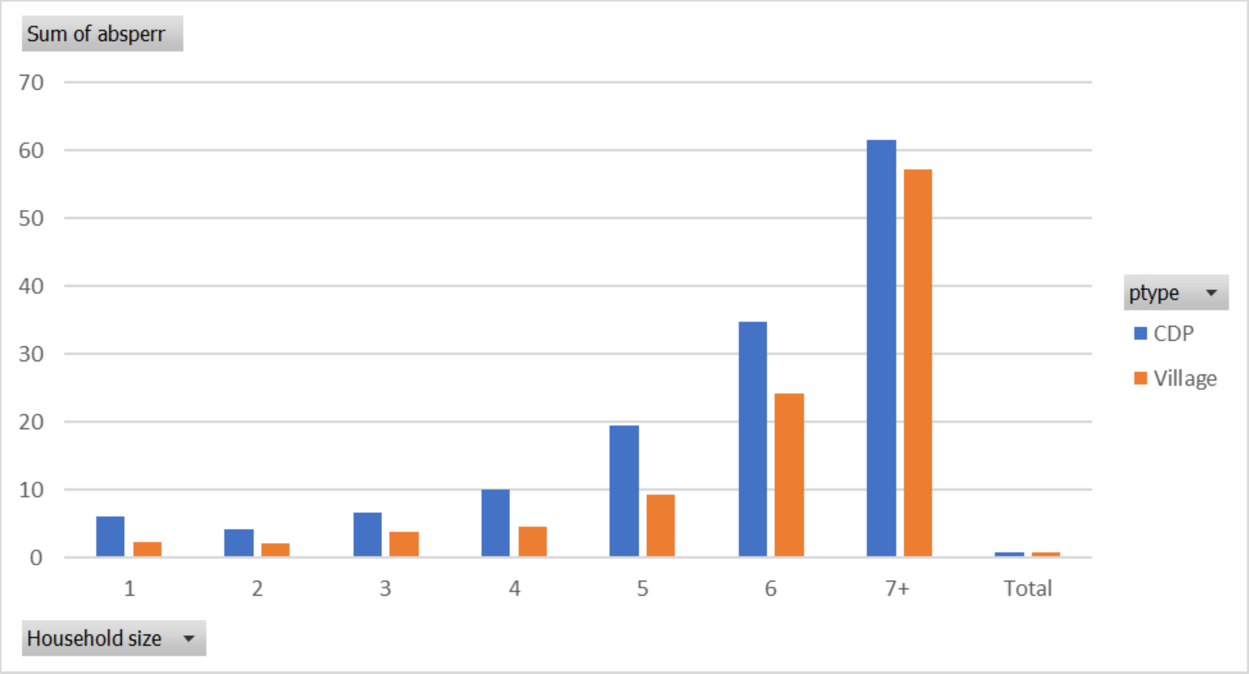


Figure 21: Average absolute percentage error by household size (MAPE), by village/CDP



5 NON-RANDOM SAMPLE OF BLOCKS

5.1 RESEARCH QUESTION:

What happens if I aggregate a collection of blocks with a very specific characteristic? How much does the TDA disturb the statistics?

5.2 CONCLUSIONS:

The noise added to blocks with a very specific characteristic can completely mask what is observed in those blocks. The example in this section is pretty extreme, but similar masking could happen with other selections or exclusions based on characteristics of the records.

5.3 METHOD AND RESULTS:

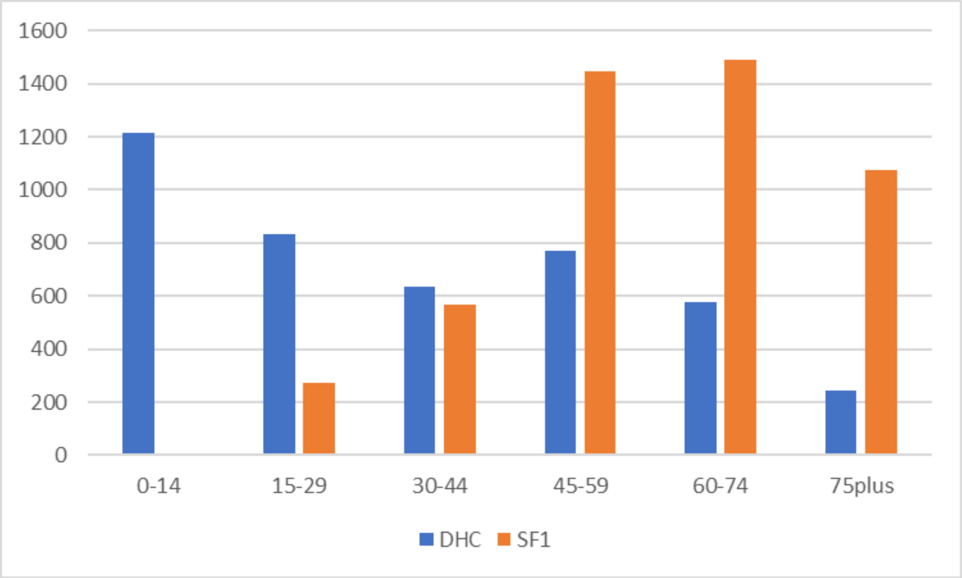
For these I analyses I looked at blocks with 1 person living in a housing unit.

In SF1 there were 4,853 such blocks in New York with 2,592 male householders living alone and 2,261 female householders living alone and as expected no other relationship observed.

In the demonstration data set, many different relationships to householder are observed:

Age	child/grandchild	HH alone	HH not alone	other	parent/in-law	son/daughter in-law	spouse/partner	Total
0-14	1136			77				1213
15-29	514	31	94	113		9	71	832
30-44	62	60	203	53	5	16	234	633
45-59	29	122	296	63	19		242	771
60-74	9	109	196	32	30	2	197	575
75plus	1	74	76	19	26		47	243
Total	1751	396	865	357	80	27	791	4267

Figure 22: Age distribution of persons living alone in a single housing unit in a block


1600
1400
1200
1000
800
600
400
200
0
0-14
15-29
30-44
45-59
60-74
75plus
DHC
SF1

6 DIFFERENCES BY TENURE

6.1 RESEARCH QUESTIONS:

Are certain measures in the Demonstration Data more vulnerable to large errors than others when separated by tenure majority (for example high rental areas compared with high home ownership areas)? This is chosen as an example to see if selecting geographies based on a characteristic (tenure in this example) can have unintended side effects.

6.2 CONCLUSIONS:

- Errors for large households (5+) were highest in rental majority areas; 36.4% of tracts and 35.8% of block groups had “Big” errors (MAE & MAPE >=10).
- Errors for counts of children under 18 present were also exacerbated in rental majority areas, especially for block groups. 7.1% of all tracts and 25.2% of all block groups had Big Errors (BE) on the number of households with children, but for rental majority areas 26.4% of tracts and 36.8% of block groups had BE.
- Counts of householder race were prone to errors between files, especially for non-White householders. The largest prevalence of big errors was found for Hispanic householders (32% of tracts and 22.2% of block groups).
  - At the tract level errors were worse in majority owned areas- 35.1% of tracts had big errors compared to 22.3% of rental-majority tracts.
  - At the block group level, errors for Hispanic householders were highest in rental majority block groups (30.4% with BE), compared to 17% in majority-owned areas.
- At the block group level, the share of geographies with BE for younger (18-24) and older (65-74) age groups decrease when split by sex but increase for the middle-age group (40-59), from 15% of block groups with BE for both sexes to 27% with BE for men and 26% for women.
- Errors on sex by age are amplified in rental-majority areas; for all sex by age groups (except for men and women ages 65-74), over 32% of block groups had big errors in these areas.
- Note: Error estimates for median age by race are distorted by the occurrence of 0 population counts in either file at the tract or block group level (e.g. zero Asian people in a tract in the DP-DHC and one Asian person age 90 in SF1 would result in an observed error of -90), and therefore we will not draw conclusions based on those measures.

6.3 METHODOLOGY:

- We used the 2010 Summary File 1 and 2010 DP DHC housing unit and person files at the Census Tract and Block Group levels of geography, excluding Puerto Rico. Person and unit files were merged by geocode at each level of geography into one full dataset.
- Our analysis focused on urban tracts and block groups, defined here as 200+ households per tract, and 150+ households per block group.
- Majority housing tenure was determined by calculating percent ownership in a tract or block group: [(IFF002_sf + IFF003_sf)/H8C001_sf] *100
- Geographies with <=20% ownership were classified as *rental majority* areas; geographies between 21% and 79% ownership were *mixed tenure* areas; geographies with >=80% ownership were *owner majority* areas.

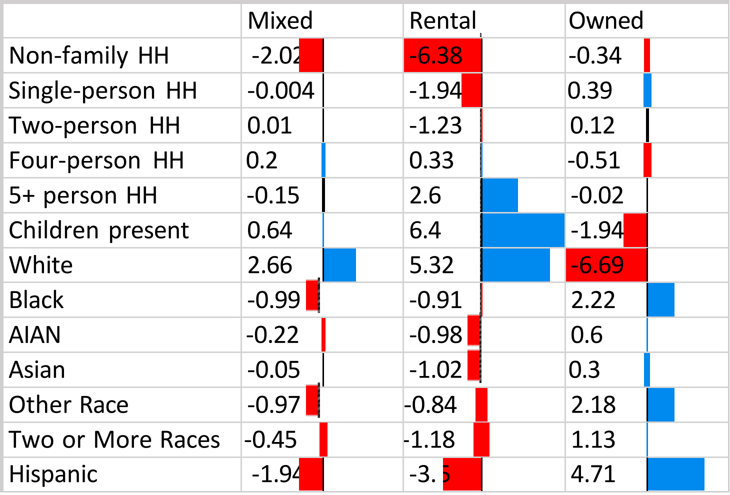
- Our tract-level dataset contained 71,842 Census tracts, and the block group dataset contained 214,558 block groups.

6.4 Metrics of Error:

Our analysis of the demonstration data in comparison with SF1 included four key accuracy metrics: Mean Error, Mean Absolute Error (MAE), Mean Percent Error (MPE), and Mean Absolute Percent Error (MAPE). We also included a measure of "Big" Errors, the sum of tracts or block groups where both the MAE & MAPE >=10.

6.5 Results:

Figure 23: Mean Error of Tract-level Unit File Variables, by Tenure Majority



	Mixed	Rental	Owned
Non-family HH	-2.02	-6.38	-0.34
Single-person HH	-0.004	-1.94	0.39
Two-person HH	0.01	-1.23	0.12
Four-person HH	0.2	0.33	-0.51
5+ person HH	-0.15	2.6	-0.02
Children present	0.64	6.4	-1.94
White	2.66	5.32	-6.69
Black	-0.99	-0.91	2.22
AIAN	-0.22	-0.98	0.6
Asian	-0.05	-1.02	0.3
Other Race	-0.97	-0.84	2.18
Two or More Races	-0.45	-1.18	1.13
Hispanic	-1.94	-3.5	4.71

Figure 24: Mean Absolute Error of Tract-level Unit File Variables, by Tenure Majority

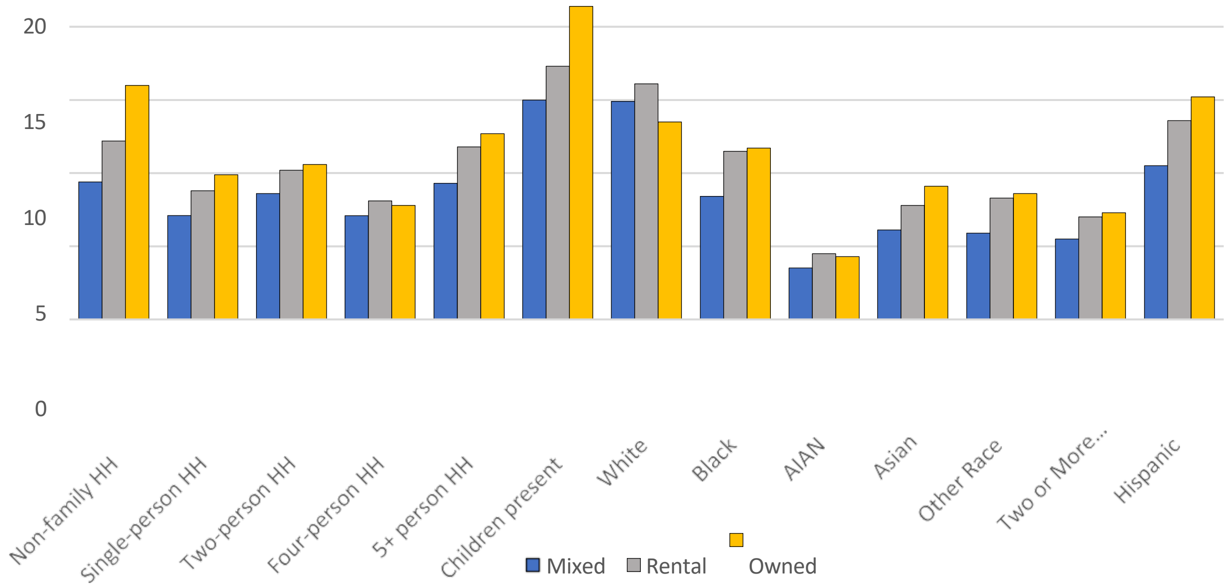


Figure 25: Mean Error of Tract-level Person File Variables, by Tenure Majority

		Mixed	Rental	Owned
Total	18-24	**-0.39**	**-3.15**	0.25
	40-59	0.16	-0.07	-0.06
	65-74	0.31	-0.27	0.03
Women	18-24	-0.16	**-1.81**	0.28
	40-59	0.09	-0.07	-0.02
	65-74	0.12	-0.16	0.01
Men	18-24	-0.23	**-1.34**	-0.03
	40-59	0.07	-0.04	-0.04
	65-74	0.19	-0.11	0.02

Figure 26: Mean Absolute Error of Tract-level Person File Variables, by Tenure Majority

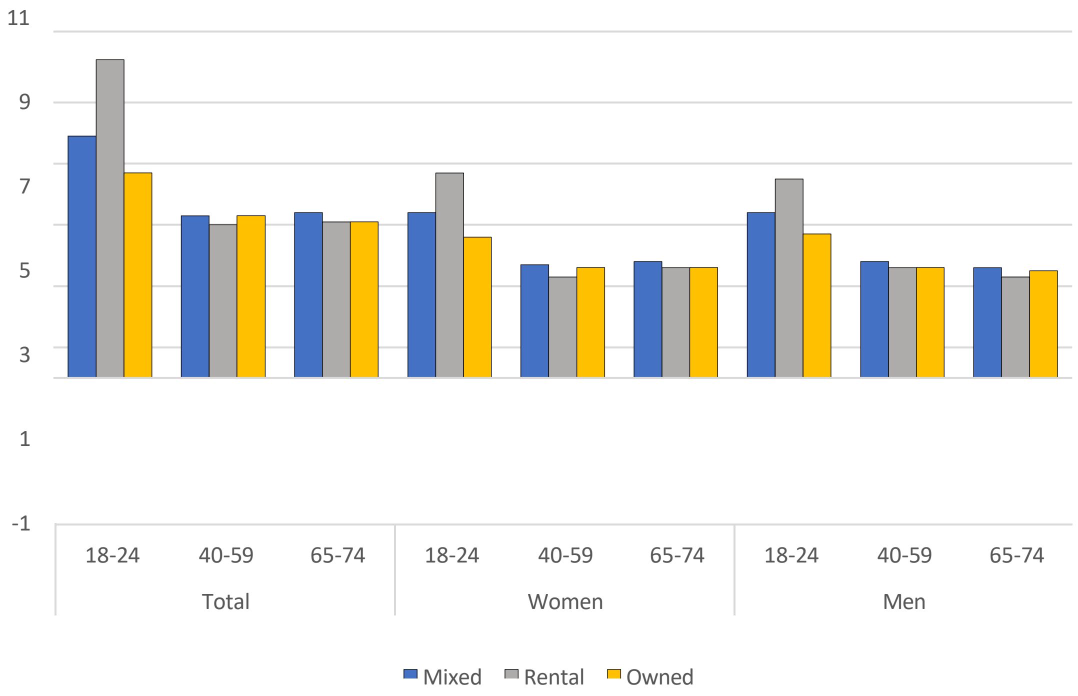


Figure 5: Mean Error of Block Group-level Unit File Variables, by Tenure Majority

Figure 27: Mean Error of Block Group-level Unit File Variables, by Tenure Majority

	Mixed	Rental	Owned
Total households	0.14	1.09	-0.42
Rental households	-0.12	-1.13	0.39
Owned households	0.26	2.22	-0.81
Non-family HH	-0.88	-3.24	0.39
Single Person HH	-0.20	-1.20	0.53
Four person HH	0.22	0.74	-0.48
5+ person HH	0.10	1.59	-0.40
Children present in H	0.35	20	-0.98
Race of Householder			
White	1.28	6.25	-3.18
Black	-0.58	-0.67	1.00
AIAN	-0.06	-0.60	0.20
Asian	0.07	-1.10	0.09
Other race	-0.49	-1.01	0.95
2+ races	-0.11	-1.49	0.44

Non-Hispanic	1.04	3.82	-2.32
Hispanic	-0.90	-2.73	1.90

Figure 28: Mean Absolute Error of Block Group-level Unit File Variables, by Tenure Majority

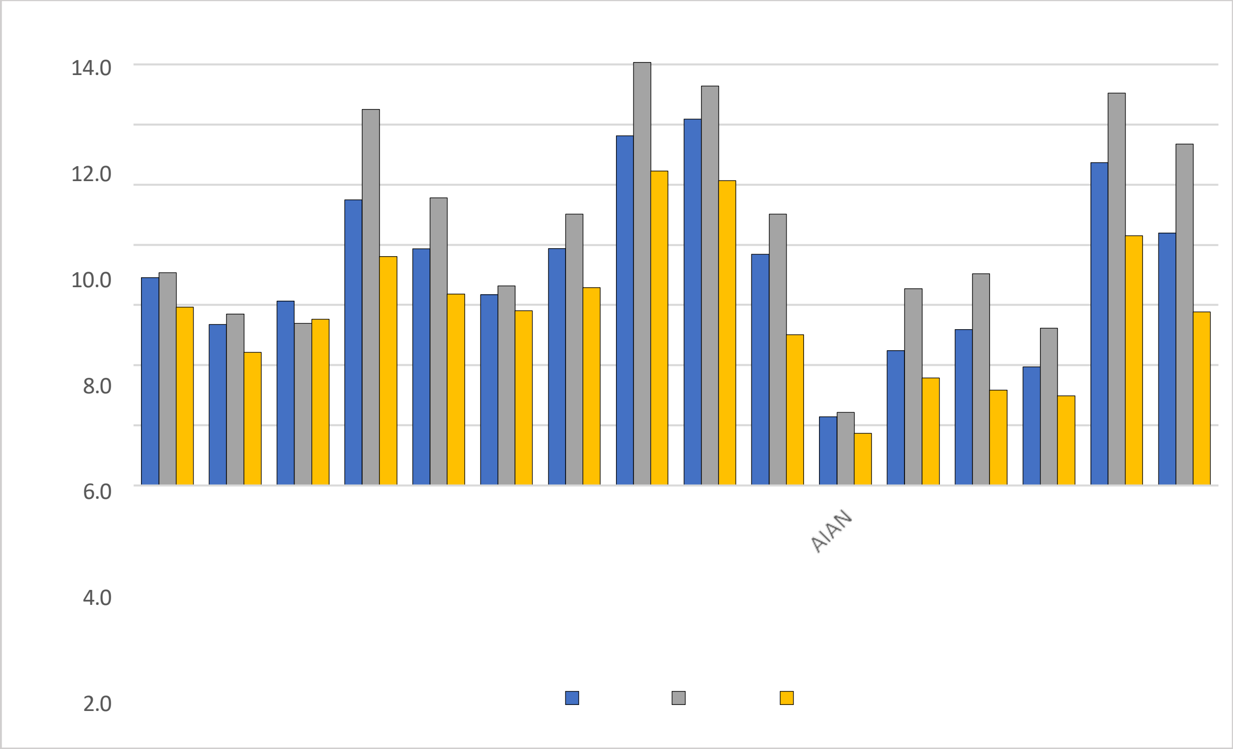
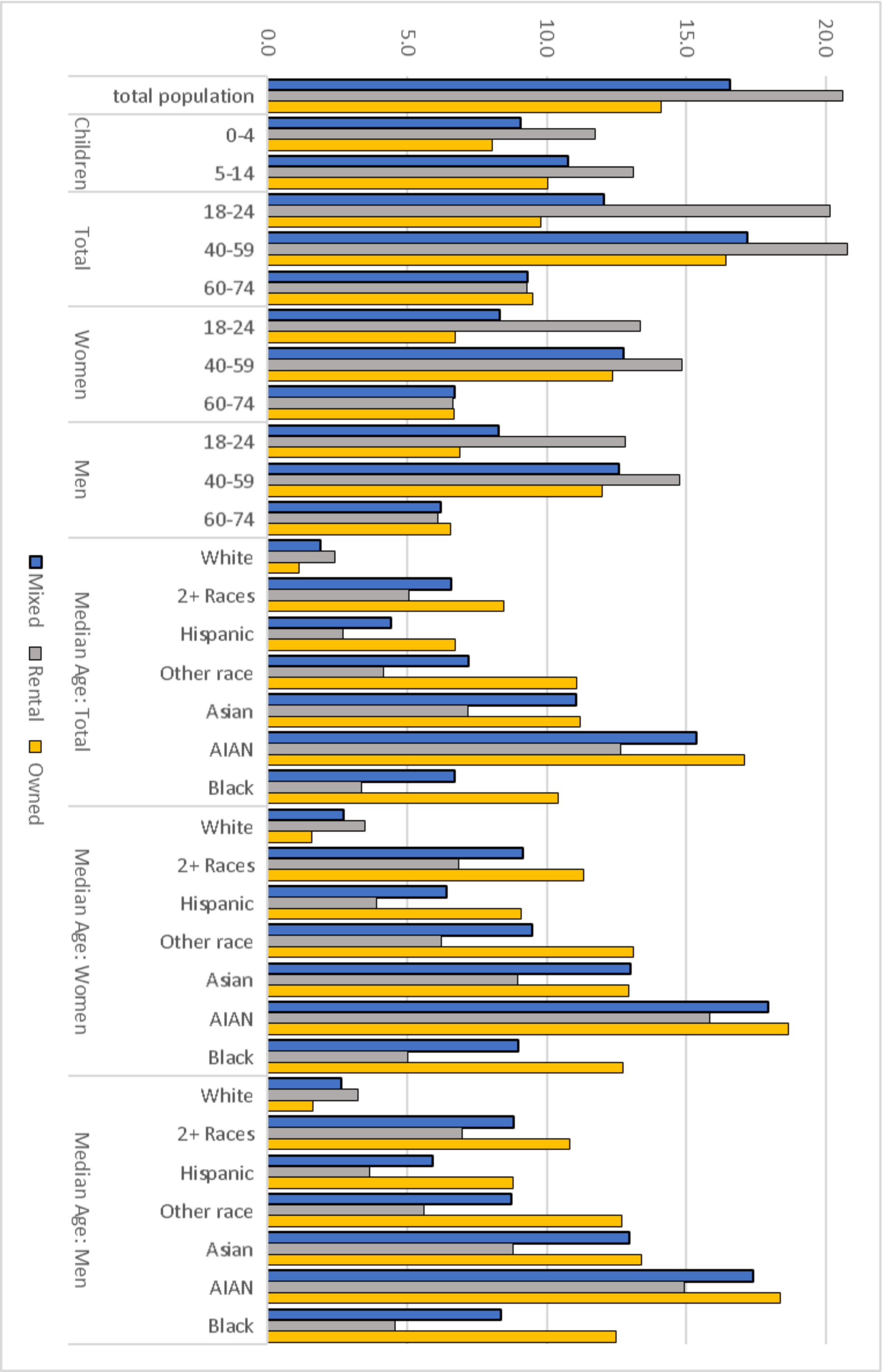


Figure 29: Mean Error of Block Group-level Person File Variables, by Tenure Majority

		Mixed	Rental	Owned
	total population	-0.4	-7.3	2.0
Children	0-4	-0.2	-1.9	0.7
	5-14	0.0	-0.4	0.1
Total	18-24	-0.1	-6.1	0.9
	40-59	0.2	2.4	-0.7
	60-74	0.2	-0.3	-0.2
Women	18-24	0.0	-3.5	0.6
	40-59	0.1	1.3	-0.4
	60-74	0.1	-0.1	-0.1
Men	18-24	0.0	-2.6	0.3
	40-59	0.1	1.0	-0.3
	60-74	0.1	-0.1	-0.1
Median Age: Total	White	0.0	0.2	-0.1
	2+ Races	-0.4	-0.5	-0.4
	Hispanic	-0.1	0.0	-0.9
	Other race	-0.2	0.1	-0.8
	Asian	-0.8	0.1	-1.8
	AIAN	-1.9	-1.6	-2.4
	Black	-0.1	0.2	-1.3
Median Age: Women	White	0.0	0.3	-0.1
	2+ Races	-0.2	-0.4	-0.2
	Hispanic	-0.1	0.1	-0.9
	Other race	-0.3	-0.1	-1.1
	Asian	-1.6	0.0	-2.7
	AIAN	-2.3	-2.0	-2.3
	Black	-0.1	0.2	-1.1
Median Age: Men	White	0.0	0.3	-0.2
	2+ Races	0.0	-0.3	-0.2
	Hispanic	0.0	0.2	-0.5
	Other race	-0.3	0.1	-0.8
	Asian	-0.3	-0.1	-1.0
	AIAN	-1.8	-1.7	-2.4
	Black	-0.3	0.3	-1.8

Figure 30: Mean Absolute Error of Block Group-level Person File Variables, by Tenure



Majority

6.6 Housing Unit File- Census Tract

Table 1: Tract-Level Prevalence of "Big" Errors Between 2010 SF1 and DP DHC

Measure (Unit/Household)	Share of Tracts with "Big" Errors
Non-family Household	3.4%
Children < 18 present	7.2%
Single-person Household	2.8%
Two-person Household	2.0%
Four-person Household	8.7%
5+ Person Household	24.2%
Race of Householder	
White	3.8%
Black	24.4%
AIAN	24.4%
Asian	21.9%
Other Race	22.0%
2 or More Races	23.4%
Hispanic	32.0%

Conclusions for Tables 2-4:

- Errors for large households (5+) were highest in rental majority areas (*Table 3*); about 36% of tracts had big errors.

- Errors for children under 18 present were worse in rental majority areas; 7% of all tracts and 26% of rental majority tracts had big errors.

-Counts of non-White householders were prone to error. The largest prevalence of Big errors was found for Hispanic householders (32% of tracts), and were larger in majority owned areas (*Table 4*) (35% BE, vs 22% BE in rental-majority tracts).

Table 2: Error Metrics of the Housing Unit File, Mixed Tenure Census Tracts

Selected Household (H) Measures by Tract-Level Housing Tenure Majority					
Mixed Tenure (*21-79% Owned***);** ***n***= **46,641**					
Measure	**Mean Error***	**Mean Absolute Error**	**Mean % Error***	**Mean Absolute % Error**	**Share of Tracts with Big Errors**
Non-family HH	-2.02*	12.2	0.1%*	2.7%	3.1%
Single-person HH	-0.004	8.8	0.2%*	2.6%	2.6%
Two-person HH	0.01	10.2	0.2%*	2.6%	2.4%
Four-person HH	0.20*	8.1	0.7%*	5.5%	10.1%
5+ person HH	-0.15*	11.8	5.7%*	12.6%	26.8%
Children present	0.64*	17.3	1.4%*	4.7%	7.1%
Householder Race/Ethnicity					
White	2.66*	16.1	5.6%*	7.2%	4.9%
Black	-0.99*	11.5	55.4%*	67.2%	24.3%
AIAN	-0.22*	4.5	52.8%*	92.8%	9.7%
Asian	-0.05	7.8	98.6%*	119.1%	24.0%
Other Race	-0.97*	8.3	40.1%*	61.5%	24.2%
Two or More Races	-0.45*	7.0	11.7%*	35.3%	26.3%
Hispanic	-1.94*	13.6	13.1%*	27.5%	31.4%

*Indicates 95% confidence that the error is statistically significantly different from 0

Table 3: Error Metrics for the Housing Unit File, Majority Rental Census Tracts

Selected Household (H) Measures by Tract-Level Housing Tenure Majority					
Majority Rental *(< 20% owned)*; ***n=3,634***					
Measure	**Mean Error***	**Mean Absolute Error**	**Mean % Error***	**Mean Absolute % Error**	**Share of Tracts With Big Errors**
Non-family HH	-6.38*	16.0	2.7%*	5.7%	6.1%
Single-person HH	-1.94*	9.9	1.1%*	3.8%	4.7%
Two-person HH	-1.23*	10.6	0.1%	3.9%	6.1%
Four-person HH	0.33	7.8	6.1%*	12.4%	15.3%
5+ person HH	2.60*	12.7	48.1%*	54.5%	36.4%
Children present	6.40*	21.4	15.2%*	18.6%	26.4%
Householder Race/Ethnicity					
White	5.32*	13.5	9.5%*	11.4%	8.0%
Black	-0.91*	11.7	5.2%*	12.2%	17.6%
AIAN	-0.98*	4.3	27.7%*	72.5%	9.4%
Asian	-1.02*	9.1	57.7%*	73.1%	24.8%
Other Race	-0.84*	8.6	21.2%*	34.2%	18.1%
Two or More Races	-1.18*	7.3	1.3%*	19.2%	26.9%
Hispanic	-3.50*	15.2	7.0%*	15.8%	22.3%

Table 4: Error Metrics for the Housing Unit File, Majority Owned Census Tracts

Selected Household (H) Measures by Tract-Level Housing Tenure Majority					
Majority Owned *(>80% owned)*; ***n=21,567***					
Measure	**Mean Error***	**Mean Absolute Error**	**Mean % Error***	**Mean Absolute % Error**	**Share of Tracts with Big Errors**
Non-family HH	-0.34*	9.4	0.3%*	3.0%	3.5%
Single-person HH	0.39*	7.1	0.3%*	2.9%	2.9%
Two-person HH	0.12	8.6	0.1%*	1.7%	0.5%
Four-person HH	-0.51*	7.1	1.2%*	4.8%	4.5%
5+ person HH	-0.02	9.3	8.7%*	14.1%	16.8%
Children present	-1.94*	15.0	6.6%*	9.6%	4.2%
Householder Race/Ethnicity					
White	-6.69*	14.9	0.04%	2.0%	0.7%
Black	2.22*	8.4	140.7%*	153.7%	25.5%
AIAN	0.60*	3.5	91.6%*	129.5%	5.6%
Asian	0.30*	6.1	126.4%*	146.4%	16.7%
Other Race	2.18*	5.9	107.4%*	127.9%	17.9%
Two or More Races	1.13*	5.5	33.1%*	56.0%	16.5%
Hispanic	4.71*	10.5	37.5%*	49.6%	35.1%

*Indicates 95% confidence that the error is statistically significantly different from 0

*Indicates 95% confidence that the error is statistically significantly different from 0

6.7 Person File- Census Tract

Table 5: Tract-Level Prevalence of "Big" Errors Between 2010 SF1 and DP DHC

Share of Tracts with "Big" Errors			
Median Age by Race	Total Pop	Women	Men
Black	11.3%	17.3%	15.1%
Asian	19.5%	23.9%	27.0%
AIAN	10.5%	48.9%	47.8%
Other Race	4.2%	20.9%	18.6%
2 or more races	8.0%	15.2%	13.9%
Hispanic	5.0%	9.4%	8.2%

Note: Measures were only displayed if the total share of "Big" errors was > 5%; "Big" error= Census tract with an absolute error >10 and an absolute percent error > 10%

Selected Population (P) Measures by Tract-Level Housing Tenure Majority						
Mixed Tenure** (21-79% Owned); **n=46,641						
		Mean Error*	**Mean Absolute Error**	**Mean % Error***	**Mean Absolute % Error**	**Share of Tracts with Big Errors**
Sex by Age						
Total	18-24	-0.39*	7.9	0.2%*	2.4%	1.5%
	40-59	0.16*	5.3	0.0%	0.6%	0.0%
	65-74	0.31*	5.4	0.1%*	2.5%	1.3%
Women	18-24	-0.16*	5.4	0.3%*	3.4%	2.7%
	40-59	0.09*	3.7	0.0%	0.8%	0.0%
	65-74	0.12*	3.8	0.1%*	3.2%	1.2%
Men	18-24	-0.23*	5.4	0.2%*	3.3%	2.5%
	40-59	0.07*	3.8	0.0%	0.8%	0.0%
	65-74	0.19*	3.6	0.2%*	3.7%	1.6%
Median Age						
Total	Black	0.11*	3.2	10.2%*	19.2%	8.3%
	AIAN	-1.33*	10.1	41.5%*	70.1%	7.6%
	Asian	-0.72*	6.1	44.8%*	61.7%	19.0%
	Other Race	-0.13*	3.7	17.6%*	30.1%	3.0%
	Two or More Races	-0.20*	3.3	1.9%*	16.4%	6.1%
	Hispanic	-0.05*	2.0	1.0%*	8.1%	3.5%
Women	Black	0.22*	4.7	34.7%*	48.1%	13.7%
	AIAN	-1.76*	13.3	136.3%*	173.7%	47.1%
	Asian	-1.12*	6.1	67.0%*	88.1%	23.8%
	Other Race	-0.11*	5.5	45.6%*	64.1%	16.7%
	Two or More Races	-0.17*	4.8	4.7%*	24.4%	12.6%
	Hispanic	-0.02	3.0	2.5%*	13.0%	7.2%
Men	Black	0.06	4.1	14.3%*	26.3%	11.3%
	AIAN	-1.34*	12.9	132.3%*	168.9%	45.9%
	Asian	-0.25*	8.1	129.3%*	151.2%	26.1%
	Other Race	-0.10*	4.8	35.0%*	51.2%	14.1%
	Two or More Races	0.08*	4.6	6.7%*	25.7%	11.7%
	Hispanic	0.02	2.7	2.7%*	12.2%	5.9%

Table 6: Error Metrics for the Person File, Mixed Tenure Census Tracts

Table 7: Error Metrics for the Person File, Majority Rental Census Tracts

Selected Population (P) Measures by Tract-Level Housing Tenure Majority						
Majority Rental** (< 20% owned)**; n=3,634						
		Mean Error	**Mean Absolute Error**	**Mean % Error**	**Mean Absolute % Error**	**Share of Tracts with Big Errors**
Sex by Age						
Total	18-24	-3.15*	10.4	0.1%	2.2%	1.4%
	40-59	-0.07	5.0	0.2%	1.3%	0.6%
	65-74	-0.27*	5.1	6.1%*	13.4%	4.5%
Women	18-24	-1.81*	6.7	0.0%	2.9%	2.0%
	40-59	-0.07	3.3	0.3%	2.0%	0.2%
	65-74	-0.16*	3.6	4.8%*	14.0%	2.2%
Men	18-24	-1.34*	6.5	0.2%*	3.0%	2.5%
	40-59	-0.04	3.6	0.2%*	1.8%	0.2%
	65-74	-0.11	3.3	6.8%*	16.5%	2.2%
Median Age						
Total	Black	-0.01	1.0	0.15%	3.1%	1.0%
	AIAN	-1.04*	7.9	11.7%*	35.6%	3.2%
	Asian	-0.15	3.9	37.2%*	47.3%	10.4%
	Other Race	-0.16	1.9	8.8%*	15.4%	1.5%
	Two or More Races	-0.56*	2.7	-0.2%	10.8%	4.4%
	Hispanic	-0.08	1.1	0.3%	4.0%	1.8%
Women	Black	0.07	1.7	0.7%*	5.3%	3.0%
	AIAN	-1.29*	11.3	111.2%*	144.3%	39.4%
	Asian	-0.30	5.0	54.2%*	67.5%	14.4%
	Other Race	-0.14	2.9	21.1%*	30.7%	7.0%
	Two or More Races	-0.40*	3.6	1.5%*	14.9%	7.9%
	Hispanic	-0.03	1.5	2.4%	3.2%	2.8%
Men	Black	-0.03	1.4	0.1%	4.5%	1.6%
	AIAN	-1.17*	10.1	89.5%*	120.7%	34.9%
	Asian	0.00	5.0	61.6%*	74.5%	14.5%
	Other Race	-0.17	2.6	24.8%*	33.9%	5.7%
	Two or More Races	-0.55*	3.6	1.0%*	16.0%	8.0%
	Hispanic	-0.09	1.4	1.9%	6.9	2.7%

*Indicates 95% confidence that the error is statistically significantly different from 0

Table 8: Error Metrics for the Person File, Majority Owned Census Tracts

Selected Population (P) Measures by Tract-Level Housing Tenure Majority						
Majority Owned *(>80% owned);* ***n=21,567***						
		Mean Error	**Mean Absolute Error**	**Mean % Error**	**Mean Absolute % Error**	**Share of Tracts with Big Errors**
Sex by Age						
Total	18-24	0.25*	6.7	0.6%*	2.4%	2.2%
	40-59	-0.06	5.3	0.0%	0.6%	0.0%
	65-74	0.03	5.1	0.0%	2.5%	0.4%
Women	18-24	0.28*	4.6	0.9%*	3.4%	3.5%
	40-59	-0.02	3.6	0.0%	0.8%	0.0%
	65-74	0.01	3.6	0.0%	3.2%	0.6%
Men	18-24	-0.03	4.7	0.6%*	3.3%	2.7%
	40-59	-0.04	3.6	0.0%	0.8%	0.0%
	65-74	0.02	3.5	0.0%	3.7%	0.6%
Median Age						
Total	Black	-0.51*	6.1	22.9%*	19.2%	19.6%
	AIAN	-2.98*	12.9	92.7%*	70.1%	17.9%
	Asian	-1.35*	6.1	28.0%*	61.7%	22.2%
	Other Race	-0.51*	7.1	46.7%*	30.1%	7.2%
	Two or More Races	0.02	4.7	6.0%*	16.4%	12.9%
	Hispanic	-0.35*	3.6	0.7%*	8.2%	8.6%
Women	Black	-0.32*	8.3	84.3%*	48.1%	27.4%
	AIAN	-3.17*	16.0	238.9%*	173.7%	54.3%
	Asian	-1.71*	8.2	40.3%*	88.1%	25.8%
	Other Race	-0.73*	9.3	88.6%*	64.1%	32.3%
	Two or More Races	0.25*	6.7	14.8%*	24.4%	22.0%
	Hispanic	-0.43*	5.1	4.5%*	13.0%	15.2%
Men	Black	-0.86*	7.7	38.0%*	26.3%	25.6%
	AIAN	-3.17*	15.9	227.3%*	168.9%	54.1%
	Asian	-0.62*	9.6	179.9%*	151.3%	31.2%
	Other Race	-0.48*	8.8	87.3%*	51.2%	30.6%
	Two or More Races	0.13	6.2	15.5%*	25.8%	19.6%
	Hispanic	-0.03	4.9	6.4%*	12.2%	14.2%

*Indicates 95% confidence that the error is statistically significantly different from 0

6.8 HOUSEHOLD/UNIT FILE- BLOCK GROUP LEVEL

Table 9: Block Group-Level Prevalence of "Big" Errors Between 2010 SF1 and DP DHC

Measure (Unit/Household)	Share of Block Groups with "Big" Errors
Children Under 18 Present	25.2%
Non-Family Household	17.5%
Single-person Household	16.1%
Two-person Household	14.3%
Four-person Household	18.8%
Household with 5 or more people	26.6%
Black Householder	17.1%
Asian Householder	10.4%
Other Race Householder	12.7%
Hispanic or Latino Householder	22.2%

Note: Measures were only displayed if the share of "Big" errors was > 10%; "Big" error= block group with an absolute error >10 and an absolute percent error > 10%

Conclusions for Tables 10-12:

- Errors for large households (5+) were highest in rental majority areas (*Table 11*)- about 36% of block groups had "Big" errors (MAE & MAPE >=10).
- Errors for counts of children under 18 were larger in rental majority areas, especially for block groups. 25% of all block groups had Big Errors (BE) on the number of households with children, but for rental majority areas about 37% of block groups had BE.
- Counts of householder race were prone to errors between files, especially for non-White householders. The largest prevalence of big errors was found for Hispanic householders (22% of block groups).
- Errors for Hispanic householders were highest in rental majority block groups (30% had big errors) (*Table 11*), compared to 17% in majority-owned areas (*Table 12*).

Table 10: Error Metrics for the Housing Unit File, Mixed Tenure Block Groups

Selected Household (H) Measures by Block Group-Level Housing Tenure Majority					
Mixed Tenure** (21%-79% Owned Units); **n=122,591					
Measure	Mean Error	Mean Absolute Error	Mean % Error	Mean Absolute % Error	Share of Block Groups with Big Errors
Non-family HH	-0.88*	9.5	0.5%*	6.4%	16.8%
Single-person HH	-0.19*	7.9	0.8%*	6.9%	16.3%
Two-person HH	-0.01	8.4	0.5%*	6.4%	16.8%
Four-person HH	0.22*	3.4	2.3%*	13.2%	20.5%
5+ person HH	0.09	7.9	10.4%*	24.2%	28.6%
Children under 18 Present	0.34*	11.7	3.1%*	9.9%	27.1%
Householder race/ethnicity					
White	1.30*	12.2	9.3%*	13.4%	10.3%
Black	-0.58*	7.7	63.1%*	84.1%	19.4%
AIAN	-0.07	2.3	65.7%*	117.5%	2.0%
Asian	0.07	4.5	103.5%*	134.8%	11.2%
Other Race	-0.50*	5.2	53.5%*	85.7%	15.5%
Two or More Races	-0.11*	4.0	31.8%*	69.2%	7.7%
Hispanic	-0.91*	8.4	27.3%*	51.0%	24.5%
Non-Hispanic	1.05*	10.8	1.6%*	4.3%	5.9%

*Indicates 95% confidence that the error is statistically significantly different from 0

Table 11: Error Metrics for the Housing Unit File, Majority Rental Block Groups

Selected Household (H) Measures by Block Group-Level Housing Tenure Majority					
Majority Rental** (< 20% owned)**; n=14,061					
Measure	Mean Error	Mean Absolute Error	Mean % Error	Mean Absolute % Error	Share of Block Groups with Big Errors
Non-family HH	-3.33*	12.8	1.8%*	7.7%	15.5%
Single-person HH	-1.20*	9.8	1.8%*	7.7%	15.2%
Two-person HH	-0.45*	9.4	0.7%*	7.9%	21.8%
Four-person HH	0.77*	6.8	10.9%*	22.8%	24.0%
5+ person HH	1.63*	9.3	51.3%*	64.4%	35.8%
Children under 18 Present	3.30*	14.5	21.4%*	27.4%	36.8%
Householder race/ethnicity					
White	6.56*	13.7	10.5%*	14.8%	17.9%
Black	-0.68*	9.3	9.5%*	24.5%	21.8%
AIAN	-0.64*	2.5	32.7%*	93.5%	2.0%
Asian	-1.15*	6.7	54.3%*	81.7%	20.3%
Other Race	-1.06*	7.3	20.9%*	46.0%	22.7%
Two or More Races	-1.58*	5.4	-0.32%	36.0%	16.5%
Hispanic	-2.89*	11.7	7.4%*	25.5%	30.4%
Non-Hispanic	4.06*	13.4	3.5%*	6.4%	11.2%

*Indicates 95% confidence that the error is statistically significantly different from 0

Table 12: Error Metrics for the Housing Unit File, Majority Owned Block Groups

Selected Household (H) Measures by Block Group-Level Housing Tenure Majority					
Majority Owned (>80% owned); n=77,906					
Measure	Mean Error	Mean Absolute Error	Mean % Error	Mean Absolute % Error	Share of Block Groups with Big Errors
Non-family HH	0.39*	7.7	1.3%*	7.6%	18.8%
Single-person HH	0.54*	6.5	1.4%*	7.9%	16.0%
Two-person HH	0.07	7.5	0.2%*	4.5%	9.0%
Four-person HH	-0.49*	5.9	2.4%*	11.4%	15.1%
5+ person HH	-0.42*	6.7	9.4%*	22.0%	21.8%
Children under 18 Present	-1.01*	10.6	9.8%*	16.1%	20.2%
Householder race/ethnicity					
White	-3.24*	10.3	0.5%*	4.0%	2.8%
Black	1.02*	5.1	119.6%*	142.3%	12.5%
AIAN	0.21*	1.8	79.9%*	127.7%	1.1%
Asian	0.09*	3.6	90.3%*	122.2%	7.4%
Other Race	0.96*	3.2	105.0%*	137.0%	6.4%
Two or More Races	0.45*	3.0	61.0%*	99.3%	3.5%
Hispanic	1.93*	5.9	56.5%*	79.2%	17.0%
Non-Hispanic	-2.36*	8.4	-0.3%*	2.2%	1.3%

*Indicates 95% confidence that the error is statistically significantly different from 0

6.9 PERSON FILE- BLOCK GROUP LEVEL

Table 13: Block Group-Level Prevalence of "Big" Errors Between 2010 SF1 and DP DHC

Measure	Subgroup	% Of Block Groups with "Big" Errors
Children	Ages 0-4	29.2%
	Ages 5-14	22.0%
Age by Sex	All 18-24	34.2%
	All 40-59	15.3%
	All 65-74	31.8%
	Women 18-24	28.8%
	Women 40-59	26.1%
	Women 65-74	23.5%
	Men 18-24	28.4%
	Men 40-59	26.9%
	Men 65-74	21.8%
Median Age: All	Black	25.2%
	AIAN	20.8%
	Asian	35.0%
	Multiracial (2+ Races)	23.2%
Median Age: Women	Black Women	32.4%
	Other Race Women	35.4%
	AIAN Women	56.2%
	Asian Women	40.2%
	Multiracial Women	34.7%
	Hispanic Women	23.5%
Median Age: Men	Black Men	31.8%
	Other Race Men	33.4%
	AIAN Men	55.5%
	Asian Men	41.5%
	Multiracial Men	33.0%
	Hispanic men	22.2%

Note: Only measures with over 20% Big errors shown

Conclusions for Tables 14-16:

- The share of geographies with Big Errors for younger (18-24) and older (65-74) age groups decreased when split by sex but increased for the middle-age group (40-59); 15% of block groups had BE for this age group but rose to 27% of block groups with big error for men and 26% for women.

- Errors on sex by age were amplified in rental-majority areas (*Table 15*). Over 32% of block groups had big errors in all selected age and age-by-sex groups (with the exception of men and women ages 65-74).

Table 14: Error Metrics for the Person File, Mixed Tenure Block Groups

Selected Population (P) Measures by Block Group-Level Housing Tenure Majority						
		Mixed Tenure *(21%-79% Owned)****; n=122,591***				
	Measure	Mean Error	Mean Absolute Error	Mean % Error	Mean Absolute % Error	Share of Block Groups with Big Errors
Age by Sex	All 18-24	-0.06	12.1	1.3%*	11.4%	34.7%
	All 40-59	0.24*	17.2	0.2%*	5.7%	16.7%
	All 65-74	0.24*	9.3	1.2%*	13.2%	32.9%
	Women 18-24	-0.05	8.3	1.8%*	16.1%	29.7%
	Women 40-59	-0.13	12.8	0.2%*	8.3%	28.2%
	Women 65-74	0.10	6.7	1.5%*	17.5%	23.6%
	Men 18-24	-0.01	8.3	1.7%*	15.9%	29.6%
	Men 40-59	0.11	12.6	0.3%*	8.6%	29.3%
	Men 65-74	0.14*	6.2	2.0%*	19.9%	21.1%
Median Age: All	Black	-0.13*	6.7	76.0%*	94.6%	20.6%
	Other	-0.17*	7.2	142.4%*	164.9%	9.7%
	AIAN	-1.90*	15.3	387.3%*	428.2%	17.1%
	Asian	-0.78*	11.0	289.8%*	317.8%	35.5%
	Multiracial (2+ Races)	-0.39*	6.5	19.0%*	44.3%	20.6%
	Hispanic	-0.15*	4.4	18.1%*	32.6%	11.7%
Median Age: Women	Black Women	-0.08	9.0	179.9%*	205.0%	28.4%
	Other Race Women	-0.30*	9.4	225.0%*	255.9%	31.1%
	AIAN Women	-2.26*	18.0	631.7%*	680.9%	56.4%
	Asian Women	-1.62*	13.0	335.2%*	369.8%	40.8%
	Multiracial Women	-0.22*	9.1	53.8%*	86.0%	32.1%
	Hispanic Women	-0.11	6.3	43.4%*	63.9%	19.6%
Median Age: Men	Black Men	-0.28*	8.4	110.1%*	134.3%	27.1%
	Other Race Men	-0.28*	8.7	195.2%*	223.7%	28.4%
	AIAN Men	-1.86*	17.4	604.7%*	653.1%	55.5%
	Asian Men	-0.35*	13.0	467.6%*	501.9%	41.6%
	Multiracial Men	-0.02	8.7	55.4%*	87.8%	30.9%
	Hispanic men	0.00	5.9	40.6%*	59.8%	17.9%

*Indicates 95% confidence that the error is statistically significantly different from 0

Table 15: Error Metrics for the Person File, Majority Rental Block Groups

Selected Population (P) Measures by Block Group-Level Housing Tenure Majority						
Majority Rental** (< 20% owned)**; n=14,061						
	Measure	Mean Error	Mean Absolute Error	Mean % Error	Mean Absolute % Error	Share of Block Groups with Big Errors
Age by Sex	All 18-24	-6.24*	20.5	-0.9%*	11.5%	38.8%
	All 40-59	2.45*	21.4	3.3%*	11.1%	32.0%
	All 65-74	-0.31*	9.6	9.2%*	31.2%	36.7%
	Women 18-24	-3.6*	13.6	-1.0%*	15.4%	40.5%
	Women 40-59	1.42*	15.3	4.1%*	15.9%	40.9%
	Women 65-74	-0.15	6.9	9.8%*	37.9%	24.5%
	Men 18-24	-2.65*	13.0	-0.1%	16.0%	40.0%
	Men 40-59	1.03*	15.3	2.9%*	15.0%	41.3%
	Men 65-74	-0.15	6.3	11.5%*	43.4%	21.9%
Median Age: All	Black	0.27*	3.2	10.2%*	18.4%	7.1%
	Other	0.06	4.0	42.2%*	54.5%	3.7%
	AIAN	-1.62*	12.8	203.5%*	241.2%	7.3%
	Asian	0.09	7.1	153.3%*	170.3%	21.0%
	Multiracial (2+ Races)	-0.58	4.9	6.2%*	25.1%	12.7%
	Hispanic	0.01	2.5	3.9%*	11.7%	4.3%
Median Age: Women	Black Women	0.31*	4.9	27.5%*	40.5%	14.1%
	Other Race Women	-0.06	6.1	79.9%*	99.4%	17.6%
	AIAN Women	-2.18*	16.1	382.8%*	431.5%	53.8%
	Asian Women	0.00	4.9	188.4%*	210.4%	27.7%
	Multiracial Women	-0.60*	6.6	12.1%*	36.4%	22.2%
	Hispanic Women	0.09	3.7	10.5%*	21.8%	8.2%
Median Age: Men	Black Men	0.29*	4.5	10.5%*	22.9%	12.1%
	Other Race Men	0.05	5.5	66.7%*	84.4%	15.5%
	AIAN Men	-1.70*	15.1	390.1%*	436.9%	50.8%
	Asian Men	-0.07	8.8	198.1%*	221.1%	28.0%
	Multiracial Men	-0.42*	6.8	19.9%*	45.7%	22.9%
	Hispanic men	0.15	3.5	13.7%*	24.6%	7.7%

*Indicates 95% confidence that the error is statistically significantly different from 0

Table 16: Error Metrics for the Person File, Majority Owned Block Groups

Selected Population (P) Measures by Block Group-Level Housing Tenure Majority						
Majority Owned** (>80% owned); **n=77,906						
	Measure	Mean Error	Mean Absolute Error	Mean % Error	Mean Absolute % Error	Share of Block Groups with Big Errors
Age by Sex	All 18-24	0.93*	9.9	3.9%*	14.1%	32.6%
	All 40-59	-0.74*	16.7	-0.1%*	4.4%	10.2%
	All 65-74	-0.15*	9.6	0.5%*	9.9%	29.3%
	Women 18-24	0.63*	6.8	4.9%*	19.9%	23.8%
	Women 40-59	-0.41*	12.6	-0.1%*	6.5%	20.2%
	Women 65-74	-0.06	6.8	0.7*	13.5%	23.1%
	Men 18-24	0.30*	7.0	3.6%*	18.2%	24.5%
	Men 40-59	-0.34*	12.2	0.0%	6.7%	20.4%
	Men 65-74	-0.09	6.7	0.6%*	14.3%	22.8%
Median Age: All	Black	-1.28*	10.5	178.2%*	208.5%	35.7%
	Other	-0.81*	11.2	288.7%*	323.5%	18.4%
	AIAN	-2.43*	17.3	591.8%*	636.3%	29.1%
	Asian	-1.85*	11.3	226.9%*	258.0%	36.7%
	Multiracial (2+ Races)	-0.42*	8.5	42.3%*	71.7%	29.1%
	Hispanic	-0.93*	6.8	25.4%*	48.1%	22.4%
Median Age: Women	Black Women	-1.13*	12.9	352.4%*	388.8%	42.0%
	Other Race Women	-1.12*	13.3	396.6%*	439.3%	45.5%
	AIAN Women	-2.40*	18.9	819.6%*	869.3%	56.5%
	Asian Women	-2.76*	13.1	256.7%*	294.1%	41.5%
	Multiracial Women	-0.25*	11.4	109.6%*	146.7%	41.1%
	Hispanic Women	-0.97*	9.1	69.0%*	98.1%	32.4%
Median Age: Men	Black Men	-1.88*	12.7	239.4%*	277.2%	42.7%
	Other Race Men	-0.82*	12.9	390.7%*	431.9%	44.4%
	AIAN Men	-2.47*	18.6	772.4%*	822.4%	56.3%
	Asian Men	-1.03*	13.6	481.2%*	518.1%	43.7%
	Multiracial Men	-0.18*	10.9	110.1%*	146.9%	38.2%
	Hispanic men	-0.47*	8.9	65.6%*	93.8%	31.5%

*Indicates 95% confidence that the error is statistically significantly different from 0

7. Erica Maurer, NYC Dept. of City Planning

On behalf of the New York City Department of City Planning (DCP), I am pleased to respond to the March 16, 2022 request for feedback on the 2020 Census Data Products.

We have limited our assessment of the March 2022 DHC demonstration data to variables that are really critical to our operations. The most important component of the DHC, from our perspective, is the 5-year age sex breakdown, because this is the key input for our population projections and estimates. We evaluated the 5-year age sex data by census tract and for our geographic unit of analysis, Neighborhood Tabulation Areas, or NTAs, which are rough approximations of New York City neighborhoods built out of census tracts. Our finding is that while the demonstration files released in October 2019 and May 2020 could be wildly inaccurate, the latest release is indeed fit for use.

Reiterating our feedback in response to the 2020 Census Data Product Planning Crosswalk, unfortunately there are still vast amounts of 2020 Census data that remain unpublished in the redistricting and DHC data. Regarding content, our primary concern is the availability of 5-year age-sex data for the household population, which under the current plan will only be offered at the county-level as part of the detailed DHC data, unlike in 2010 when tract-level data were available. These data are another essential input for the preparation of population projections at the neighborhood-level using the cohort component model. Lacking such inputs, New York City will lose the precision we rely upon to direct billions of dollars in resources – resources directed towards a host of endeavors, from new school construction to the siting of our elder care facilities, essential elements for planning the future of our city.

Another great concern is the reduced geographic specificity associated with detailed race and Hispanic Origin data – the current proposal offers detail down to the county-level, whereas the 2010 Census had detail down to a census tract-level, which is crucial to our understanding of the nuance underlying race and Hispanic ethnicity. In New York City, it is not enough to know, for example, that the Asian population has decreased in Manhattan’s Chinatown. We must disentangle subgroup information by race, distinguishing whether it was the Chinese or Vietnamese population that declined in this example, so that we can properly allocate resources for services that our residents require. It is important to consider that even when an overall race group remains unchanged, we may still see significant ethnic transitions among detailed racial subgroups. The Census Bureau has invested years of work towards improving and expanding the race and Hispanic questions, so that we can more precisely portray our increasingly diverse population. For the 2020 Census, the Bureau collected roughly 350 million write-in responses across all racial and Hispanic ethnicity groups, compared to about 50 million collected in 2010. However, the current product plan does not do justice to this collection effort, and respondent burden, because it fails to tap the rich responses that can better portray the diversity of neighborhoods across the nation.[1]

In summary, while the DHC data we evaluated seem fit for use, it is critical for our work that the data that were previously available at lower levels of geographies that we frequently used are made available from the 2020 Census in the same regard. Our hope is that 5-year age-sex data for the household population are part of the set of tables slated to be reinstated in the next DHC demonstration data release. Along with that, we strongly recommend that the detailed race and Hispanic Origin data be released down to a tract-level.

Sincerely,

Erica Maurer
Senior Demographic Analyst
NYC DEPT. OF CITY PLANNING

[1] In New York City, these current proposals to reduce critical 2020 Census data detail from the census tract level up to the county-level will be particularly damaging, as our smallest county has nearly a half million people, while our average tract has a population of about 4,000. Consequently, the current proposal will reduce geographic detail for key characteristics by more than 100-fold. Unfortunately, the same can be said for many locales across the country.

May 16, 2022
RE: 2020 Census Data Products

On behalf of the New York City Department of City Planning (DCP), I am pleased to respond to the March 16, 2022 request for feedback on the *2020 Census Data Products*.

We have limited our assessment of the March 2022 DHC demonstration data to variables that are really critical to our operations. The most important component of the DHC, from our perspective, is the 5-year age sex breakdown, because this is the key input for our population projections and estimates. We evaluated the 5-year age sex data by census tract and for our geographic unit of analysis, Neighborhood Tabulation Areas, or NTAs, which are rough approximations of New York City neighborhoods built out of census tracts. Our finding is that while the demonstration files released in October 2019 and May 2020 could be wildly inaccurate, the latest release is indeed fit for use.

Reiterating our feedback in response to the 2020 Census Data Product Planning Crosswalk, unfortunately there are still vast amounts of 2020 Census data that remain unpublished in the redistricting and DHC data. Regarding content, our primary concern is the availability of 5-year age-sex data for the household population, which under the current plan will only be offered at the county-level as part of the detailed DHC data, unlike in 2010 when tract-level data were available. These data are another essential input for the preparation of population projections at the neighborhood-level using the cohort component model. Lacking such inputs, New York City will lose the precision we rely upon to direct billions of dollars in resources – resources directed towards a host of endeavors, from new school construction to the siting of our elder care facilities, essential elements for planning the future of our city.

Another great concern is the reduced geographic specificity associated with detailed race and Hispanic Origin data – the current proposal offers detail down to the county-level, whereas the 2010 Census had detail down to a census tract-level, which is crucial to our understanding of the nuance underlying race and Hispanic ethnicity. In New York City, it is not enough to know, for example, that the Asian population has decreased in Manhattan’s Chinatown. We must disentangle subgroup information by race, distinguishing whether it was the Chinese or Vietnamese population that declined in this example, so that we can properly allocate resources for services that our residents require. It is important to consider that even when an overall race group remains unchanged, we may still see significant ethnic transitions among detailed racial subgroups. The Census Bureau has invested years of work towards improving and expanding the race and Hispanic questions, so that we can more precisely portray our increasingly diverse population. For the 2020 Census, the Bureau collected roughly 350 million write-in responses across all racial and Hispanic ethnicity groups, compared to about 50 million collected in 2010. However, the current product plan does not do justice to this collection effort, and respondent burden, because it fails to tap the rich responses that can better portray the diversity of neighborhoods across the nation.[1]

[1] In New York City, these current proposals to reduce critical 2020 Census data detail from the census tract level up to the county-level will be particularly damaging, as our smallest county has nearly a half million people, while our average tract has a population of about 4,000. Consequently, the current proposal will reduce geographic detail for key characteristics by more than 100-fold. Unfortunately, the same can be said for many locales across the country.

In summary, while the DHC data we evaluated seem fit for use, it is critical for our work that the data that were previously available at lower levels of geographies that we frequently used are made available from the 2020 Census in the same regard. Our hope is that 5-year age-sex data for the household population are part of the set of tables slated to be reinstated in the next DHC demonstration data release. Along with that, we strongly recommend that the detailed race and Hispanic Origin data be released down to a tract-level.

Sincerely,
Erica Maurer
Senior Demographic Analyst
Population Division

NYC DEPT. OF CITY PLANNING

120 BROADWAY, 31st FLOOR • NEW YORK, NY 10271

www.nyc.gov/population

8. Lester Jones, National Beer Wholesalers Association

[Excerpt from email correspondence]

[For t]he alcohol industry, knowing the share of 21 and older population is critical to the adverting and marketing efforts of the industry. This basic statistic drives media spending and advertising. (See: Beer Institute Advertising/Marketing Code and Buying Guidelines)

9. Demographer (Name Withheld)

May 16, 2022

Rob Santos, Director
U.S. Census Bureau
4600 Silver Hill Road, Room 8H001
Washington DC 20233

Dear Director Santos,

I do not understand the current rational for eliminating in the proposed Demographic and Housing Characteristics file the PCT12A SEX BY AGE (WHITE ALONE) through PCT12O SEX BY AGE (TWO OR MORE RACES, NOT HISPANIC OR LATINO) tables. I do not understand what tradeoffs by population detail and geography were made to do this. When I was Chair of the Federal State Cooperative for Population Estimates we repeatedly asked for these tables to be preserved at the county level. At the 2019 workshop on data products, I highlighted in my presentation that at least 12 states at that time use these tables in their projections. I also wrote about this when comments were being solicited on the draft crosswalk table in December. I have attached that correspondence as well.

There is currently a proposed table, PCT1, that would have singe year of age for the total population at the tract level and that is the same as the PCT12 table from 2010 for the total population.

There is the P12 tables which break down the following cohorts by race and ethnicity by block. I would propose consolidating some of these groups to reflect categories such as preschool, school age, college age, prime working force, early retires, and retires or senior by block group or tract as the lowest level of geography. I only suggest block group because there may be some user that needs that level of detail, but it is likely few and far between.

Under 5 years	preschool
5 to 9 years	school age
10 to 14 years	
15 to 17 years	
18 and 19 years	college age
20 years	
21 years	
22 to 24 years	
25 to 29 years	, prime working force
30 to 34 years	
35 to 39 years	
40 to 44 years	
45 to 49 years	
50 to 54 years	
55 to 59 years	early retires
60 and 61 years	
62 to 64 years	
65 and 66 years	
67 to 69 years	
70 to 74 years	
75 to 79 years	
80 to 84 years	
85 years and over	

I will again repeat that the PCT12A SEX BY AGE (WHITE ALONE) through PCT12O SEX BY AGE (TWO OR MORE RACES, NOT HISPANIC OR LATINO) tables are needed by at least 12 states at the county level. This has been raised with staff from the Population Division as well as directly with John Abowd and Michael Haws in meetings with them. But it seems that the Bureau really has not interest in having discussions with users about tradeoffs and at what level of characteristic and geographic data is needed.

Instead, the user community is asked to present use cases so that the differential privacy algorithm can be tuned to them. This allows the Bureau to meet the demands of the squeaky wheel and does not preserve the data as a national resource.

What is sad about this is that much of what has been corrected in the PL data and the DHC for the demonstration files has been a result of the work by Jan Vink for Cornell and David Van Riper for IPUMS.

Sincerely,

Demographer (Name Withheld)

October 22, 2021

Ron Jarmin, Acting Director
U.S. Census Bureau
4600 Silver Hill Road, Room 8H001
Washington DC 20233

Dear Acting Director Jarmin,

I am disappointed to see that the PCT12A SEX BY AGE (WHITE ALONE) through PCT12O SEX BY AGE (TWO OR MORE RACES, NOT HISPANIC OR LATINO) tables are still not included in the planned DHC tables for any level of geography. On the other hand, the PCT12 SEX BY SINGLE-YEAR-AGE table is planned to be released at the tract level.

These tables are important to more than a dozen states that do their own age, sex, race, and Hispanic origin estimates and projections. This data is important at the county level for conducting this work. It seems that if 2010 data was available by block or tract it was deleted for 2020 without considering whether to keep it at the county or place level.

For example, a large number of tables on family composition and household size are entirely deleted. Many of them are needed at the place or county level to help understand the impact of COVID on housing and housing costs. At best, it appears that the decision for deleting these tables was solely based on whether it was available at the block or tract level in 2010.

There are a number of approaches for how to look at what data the Census Bureau should be providing for making well informed decisions regarding the distribution of public resources and policies for the nation's health, safety, and welfare. Census data is a national resource and the Bureau's current dissemination proposal is disconnected from that fact.

One approach is to prune back the current tables. That is, look at the history of why the various SF1 and SF2 tables were developed over the decades. What Census Bureau internal user requested the data? What outside users requested the tables? Are those tables still needed and if so, at what geographic level for helping improve the lives of the American people?

Another approach is to treat the Nation's data needs as a blank slate. This requires looking at what characteristic and geography detail data is needed at for making good decisions. The Bureau can achieve this by moving past its current placation of well-known user groups to a partnership with groups that best capture the breadth and depth of data needs for informing health, education, and transportation policies. (Please see the attached Ladder of Citizen Participation article).

I have attached proposals by the Massive Data Project and the City of New York that offer approaches for making more accurate data available for users than the Bureau's current approach allows. Both proposals show that the user community is more than willing to look at the trade-off between collapsing categories, especially the numerous race and ethnicity categories, and geographic detail to help meet the needs data for local and state entities from the public and private sectors. Again, this means that the Bureau needs to fully engage users as partners to ensure accurate data.

The transcript of the Planning for Upcoming 2020 Census Data Products September 30, 2021, online seminar states that "there was a pretty robust analysis that had been done inside the Census Bureau, things like what our internal data users made use of most from the decennial data products, what tables people downloaded most from our external facing web sites."   What is left out of this statement is that the 2010 data was in the public domain through DVD products. Also, it does not account for data, such as the PCT12 table and its iterations, being available through other agencies, private vendors, or other ways the 2010 data was in the public domain. A table could be downloaded once but that does not account for how often it might be republished through various documents. It takes time and collaborative effort for users to fully understand the strengths and weaknesses of the decennial censuses.

Users are beginning to review the 2020 Census data and they are raising concerns about the quality of the data at the national, state, and local level. These concerns include differential undercounts by age, race, and ethnicity; incorrectly geocoded group quarters; and the undercounting of housing units in rural areas. Differential privacy complicates any review of the data.

Questions about data quality combined with the planned deletion of tables will make it harder to understand what happened to country because of COVID.

I hope you will consider ways to fully partner with users for ensuring the country's data needs. The current proposal and the process of just accepting feedback on that is insufficient for finding a way to provide the most accurate data possible that is useful at needed geographic levels.

Sincerely,

Demographer (Name Withheld)

10. Angela Werner, National Center for Environmental Health

TO:

ron.s.jarmin@census.gov;
christa.d.jones@census.gov;
john.maron.abowd@census.gov;
karen.battle@census.gov;
michael.b.hawes@census.gov;
victoria.a.velkoff@census.gov

CC:

Moyer, Brian (CDC/DDPHSS/NCHS/OD) <qbk2@cdc.gov>;
Werner, Angela (CDC/DDNID/NCEH/DEHSP) <myo6@cdc.gov>;
Bunnell, Rebecca (CDC/DDPHSS/OS/OD) <rrb7@cdc.gov>;
Layden, Jennifer (CDC/DDPHSS/OS/OD) <qbg5@cdc.gov>;
Williamson, G. David (CDC/DDNID/NCEH/OD) <dxw2@cdc.gov>;
Jernigan, Daniel B. (CDC/DDPHSS/OD) <dbj0@cdc.gov>;

Cono, Joanne (CDC/DDPHSS/OS/OD) <bzc6@cdc.gov>

Dear Ron,

I hope this note finds you well.

Thank you and your colleagues for meeting with us to discuss CDC's concerns regarding Differential Privacy and the 2020 Decennial Census. As you recommended, we have re-analyzed our Impact Statements based on the newly available, March 2022 demonstration data. Below, you will find our most recent packet of Impact Statements for your review.

Below is a summary of our key findings, based on the updated Statements:

- Looking at the total population counts, there is general improvement to the data in some areas (Alaska regional areas or rural villages, for example), except for those areas with very small populations. It is worth noting that total population counts may not be the most helpful metric; calculated rates may be preferred, as they are often more responsive to differences in population distributions.
- County-level data show some overall improvement when calculating age-adjusted rates (no stratification by sex or by race/ethnicity). There are still, however, significant differences in rates using the updated data, particularly in counties with smaller populations and when stratifying the age-adjusted rates (for example, the COVID-19, age-adjusted rates by race/ethnicity and by rural/urban areas).
- County-level data remain problematic when estimating age-specific rates, including larger populations of up to 10,000 people.
- Census tract-level data remain problematic when calculating age-adjusted rates. Total population counts may not change significantly, but population changes within individual age groups can significantly impact the overall age-adjusted rate calculations.
- Observations about block-level data:
    - Used by CDC for emergency response purposes to do environmental assessments when working with communities near environmental sites, and other analyses.

- Noted the Census Bureau’s view that block-level data will not be reliable.
- Noted the variation in estimates, regardless of population density.
- CDC will not be able to accurately characterize risks and identify/target vulnerable populations using block-level data.

- Block-level maps will be unreliable, regardless of any aggregation of the block-level data.

We welcome a follow-up discussion and your continued support and collaboration as we explore the best options for using the 2020 Decennial Census data across CDC programs.

Again, thank you very much, and please do not hesitate to reach out.

Warm regards,

Brian C. Moyer, Ph.D.

Director, National Center for Health Statistics
Centers for Disease Control and Prevention

Table of Contents

CDC COVID-19 Response

Title of project: Assessing the impact of differential privacy on the age-adjusted incidence of COVID-19 at the county level

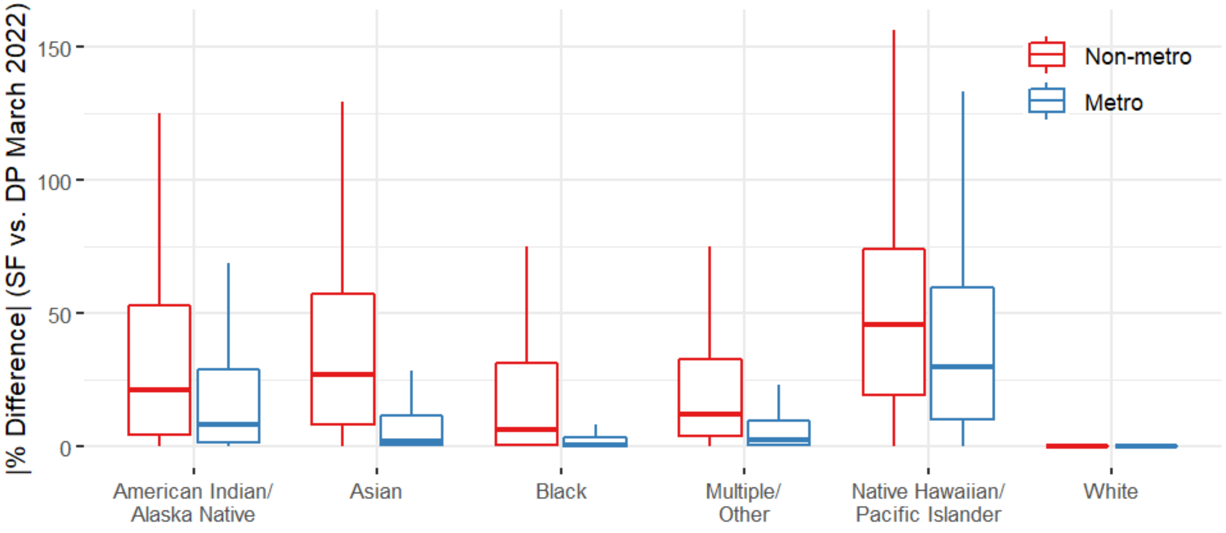
CIO/Division/Program: National Center for Environmental Health/Division of Environmental Health Science and Practice/National Environmental Public Health Tracking Program (facilitated through data collected by the Case Data Section, Data, Analytics, and Visualization Task Force)

Project description: Differential privacy is a statistical adjustment of population counts in public use datasets to protect the privacy of respondents from unauthorized disclosure. When those population counts are used as denominators for computing COVID-19 incidence rates, they may differ from the true values because of the differential privacy adjustment. To facilitate assessment of the expected impact of differential privacy, the Census Bureau released a differentially private version of the 2010 Census data. We used the most recent version of the demonstration dataset (v3-16-2022) to assess the impact of differential privacy on the 2020 age-adjusted incidence of COVID-19 at the county level stratified by race. We calculated age-adjusted incidence as this is a standard measure used in public health, and this allows for comparisons between populations with different age structures.

Methods: The number of COVID-19 cases that occurred in 2020 by county, race, and age were reported to CDC via the National Notifiable Disease Surveillance System (NNDSS), direct data entry (a legacy format used February – October 2020), or by the direct submission of CSV tables by state health departments. Cases were included if case report date, age, race, and county of residence were submitted. Age-adjusted incidence rates were calculated with and without differential privacy by dividing the reported number of COVID-19 cases for 18 separate age groups by the total 2010 population and differentially private 2010 population of each age group, with age standardization completed using the 2000 U.S. Standard Population. The absolute value of the percent difference between the age-adjusted incidence rates generated using the enumerated 2010 population counts and age-adjusted incidence rates generated using the differential privacy demonstration dataset were calculated.

Impact on project: 2020 COVID-19 age-adjusted incidence rates for minority groups were disproportionately affected by the implementation of differential privacy. Age-adjusted incidence rates for American Indian/Alaska Native, Asian, Black, Native Hawaiian/Pacific Islander, and multi-racial/other populations were particularly impacted by differential privacy, especially in non-metropolitan counties (Figure 1). The most significant divergence was observed in Native Hawaiian/Pacific Islander populations with a median difference in age-adjusted incidence rates of 45.7% in non-metropolitan counties and 29.9% in metropolitan counties (Figure 1). Age-adjusted incidence rates among Black populations were highly affected across a broad geographic area, most notably in counties with smaller Black populations (Figure 2).

Societal impact: These results demonstrate the profound impact differential privacy could have on COVID-19 incidence rates by race when differentially private Census 2020 denominators are used to compute those rates. Because racial/ethnic minority populations in counties are disproportionately affected by COVID-19 and other public health threats, the use of differentially private Census 2020 population counts could artificially increase disparities in county-level COVID-19 incidence rates, affecting existing health equity challenges. Differential privacy tends to have greater impacts on smaller populations (e.g., rural populations, minority groups). Because racial or ethnic minority populations are typically smaller populations, the age-adjusted rates for these groups are more sensitive to the effects of differential privacy. The smaller population sizes exacerbated this issue, with especially skewed rates occurring in non-Hispanic American Indian/Alaska Native, non-Hispanic Asian, non-Hispanic Black, non-Hispanic Native Hawaiian/Pacific Islander, and multi-racial/other populations, particularly those populations in rural areas. These large statistical artifacts created by differential privacy can distort the agency's health equity efforts if we are unable to distinguish real increases or decreases in COVID-19 incidence from changes caused by noise injected in the Census population denominators used. In this example, this could result in improperly allocating scarce medical resources during a pandemic and incorrectly targeting or withholding resources for vaccination based on the assumption that increases in COVID-19 incidence are real or unreal.



Source: US Census Bureau 2022; CDC Environmental Public Health Tracking Program

Figure 1: Absolute value of the percent difference between 2020 COVID-19 age-adjusted incidence rates calculated with 2010 enumerated Census population counts and age-adjusted incidence rates calculated with 2010 differential privacy demonstration population counts released in March 2022 (v3-16-2022; most recent version). Red boxes include non-metropolitan counties and blue boxes include metropolitan counties. SF = Summary File 1 data file from Census, which includes data on sex, age, race. DP = Differential Privacy demonstration dataset released in March 2022 to assess the impact of differential privacy, which includes data on sex, age, race.

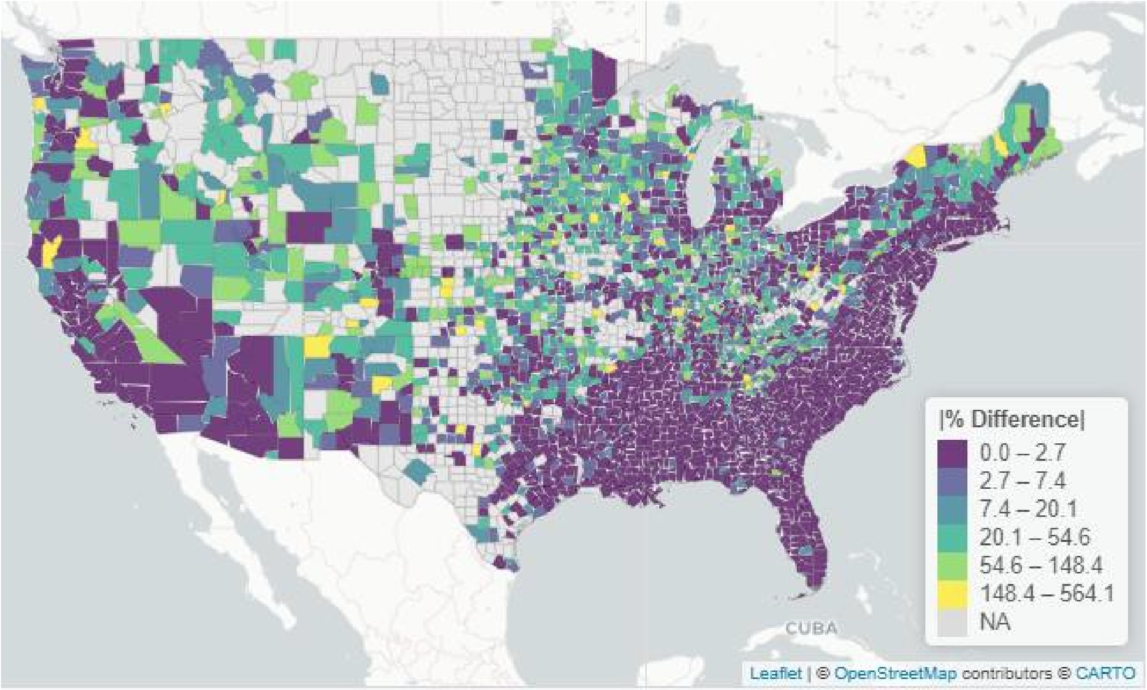


Figure 2: Absolute value of the percent difference between 2020 COVID-19 age-adjusted incidence rates for Black individuals calculated with enumerated 2010 population counts and age-adjusted incidence rates calculated with 2010 differential privacy demonstration population counts released in March 2022 (v3-16-2022; most recent version). NA values occur when no cases were reported among Black individuals in that county during 2020.

National Center for Environmental Health/National Environmental Public Health Tracking Program

Title of project: Assessing the impact of differential privacy on the incidence of health outcomes displayed on the Tracking Network (https://ephtracking.cdc.gov/DataExplorer/) at different geographic levels

CIO/Division/Program: National Center for Environmental Health/Division of Environmental Health Science and Practice/National Environmental Public Health Tracking Program

Project description: In order to facilitate assessment of the impact of differential privacy, the Census Bureau released a differentially private version (v3-16-2022) of the 2010 Census data. We used this demonstration dataset to assess the implications of differential privacy on age-adjusted rates of asthma emergency department (ED) visits and acute myocardial infarction (AMI) hospitalization at the county and census tract levels.

Methods: County- and census tract-level counts of asthma ED visits and AMI hospitalizations were acquired from recipients of the Environmental Public Health Tracking Program. In total, asthma ED data were acquired from 30 states at the county level and 6 states at the census tract level. AMI hospitalization data were acquired from 31 states at the county level and 7 states at the census tract level. Age-adjusted rates were calculated with and without differential privacy by dividing the reported number of asthma ED visits and AMI hospitalizations for 18 separate age groups by the total 2010 population and differentially private 2010 population of each age group, with age standardization completed using the 2000 U.S. Standard Population.

Impact on project: Differential privacy had minimal effects on the estimated rate of asthma ED visits at the county level (Fig. 1a). At the census tract level, changes in the rate of asthma ED visits were **generally less than 2-fold**, though differential privacy had significant effects in several census tracts, including one in which the rate of asthma ED visits **increased over 400-fold** (Fig. 1b). Only minor changes in AMI hospitalization rates were detected at the county level (Fig. 2a), though changes in the rate of hospitalizations at census tract level **routinely exceeded 5%** (Fig. 2b). While the total population count typically did not change substantially as a result of differential privacy, age-adjusted rates were sensitive to population changes within individual age groups at the census tract level.

Societal impact: Changes in population counts due to differential privacy could result in significantly overestimated (or significantly underestimated) rates, particularly at finer spatial resolutions such as census tract level. Small population sizes tended to exacerbate this issue with especially skewed rates occurring at the census tract level. This is particularly important as the Tracking Program moves to displaying and disseminating sub-county data.

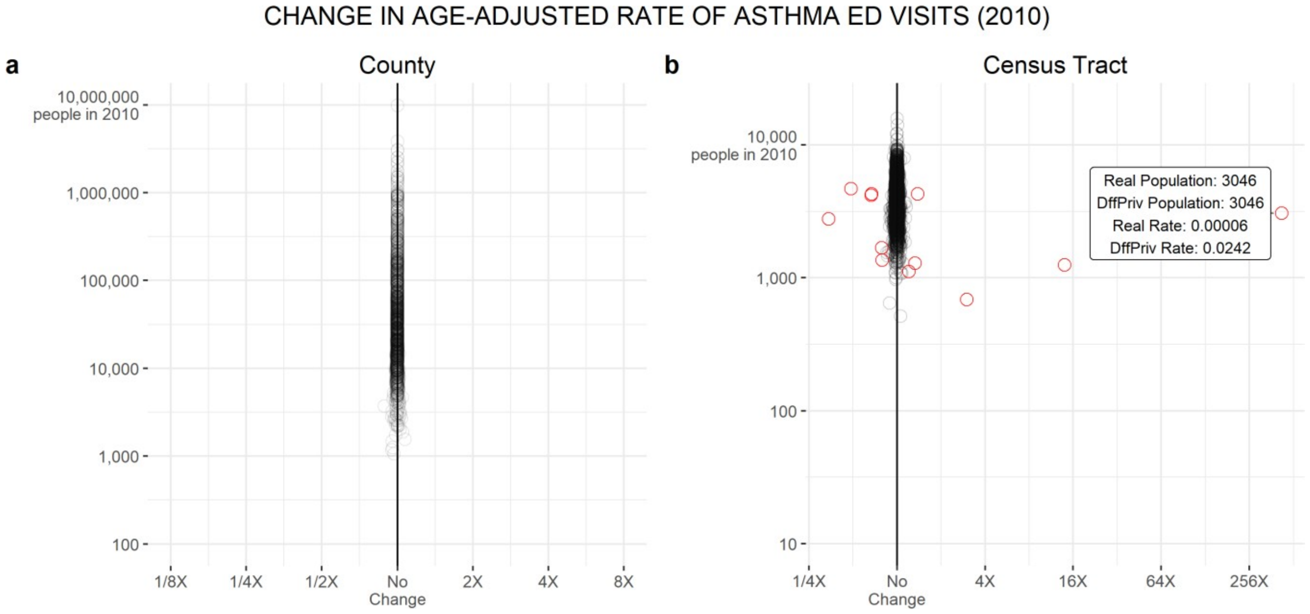


Figure 1: Change in the age-adjusted rate of asthma emergency department visits in 2010 due differential privacy population adjustments at the a) county (30 states) and b) census tract level (6 states). Counties and census tracts with at least a 20% change from the true rate are in red.

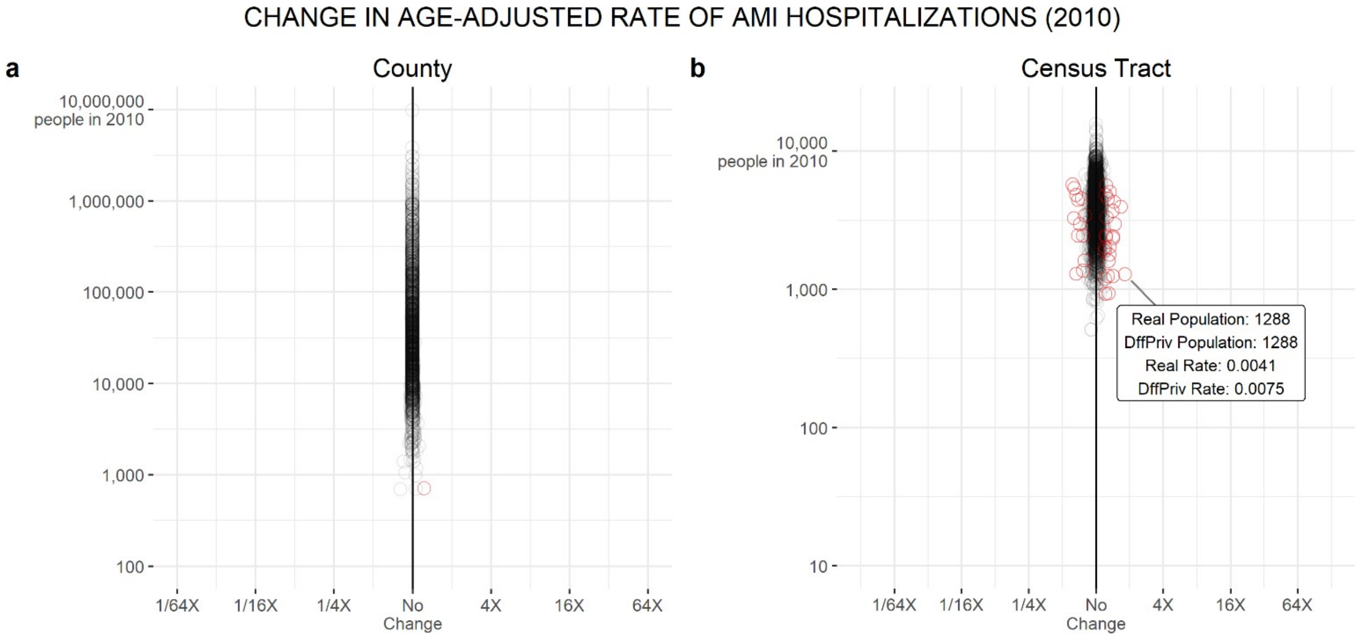


Figure 2: Change in the age-adjusted rate of acute myocardial infarction (AMI) hospitalization in 2010 due differential privacy population adjustments at the a) county (31 states) and b) census tract level (7 states). Counties and census tracts with at least a 20% change from the true rate are in red.

National Center for Chronic Disease Prevention and Health Promotion/Division for Heart Disease and Stroke Prevention

Title of project: Assessing the impact of differential privacy on estimated county-level heart disease mortality overall and by sub-group

CIO/Division/Program: National Center for Chronic Disease Prevention and Health Promotion/Division for Heart Disease and Stroke Prevention

Project description: To facilitate the assessment of the potential impact of differential privacy, the Census Bureau released a differentially private version of the 2010 Census population data. We used the most recent version (v3-16-2022) of the demonstration dataset to assess the implications of differential privacy on estimated county-level death rates for heart disease, the nation's leading cause of death.

Method: We obtained county-level heart disease death counts for the year 2010 from the National Vital Statistics System in the National Center for Health Statistics (NCHS). With these death counts, we then estimated county-level rates using two sets of denominators: (1) bridged-race populations provided by NCHS and (2) the U.S. Census Bureau's differentially private populations. To generate these estimates, we used a Bayesian spatiotemporal conditional autoregressive model that has been used extensively to examine spatiotemporal trends in cardiovascular disease death rates. Briefly, this model estimates more precise, reliable rates by incorporating correlation across space and demographic group, even in the presence of small death counts and small populations.

With this model and each population dataset, we estimated two sets of county-level rates for (1) the entire population (i.e., overall rates), and (2) stratified by both 10-year age groups and sex, resulting in four sets of rate estimates. The overall death rates were age-standardized to the 2010 U.S. population using 10-year age groups. These two sets of rates represent scenarios based on higher death counts and populations (overall rates) and based on smaller death counts and populations (rates by age group and sex). We calculated the percent change between the rates generated using the NCHS populations and the Census differential privacy populations.

Impact on project: For the overall age-standardized rates, only 0.3% of rates had more than a 20% difference between the rates estimated using differential privacy and NCHS populations (Figure 1). However, for the rates stratified by age group and sex, almost half (43.3%) of rates had more than a 20% difference between the rates estimated using differential privacy and NCHS populations. For both the overall and stratified rates, 98.0% of rates with more than a 20% difference occurred in populations of less than 10,000 people. Although some rates estimated using the differential privacy populations were lower than those estimated using the NCHS population, higher estimates were more common.

Societal impact: This analysis shows that changes in population counts resulting from the differential privacy algorithm could lead to large differences in the estimates of heart disease death rates, especially for populations less than 10,000 people. This change would severely hamper the ability to report and intervene upon heart disease in small populations, such as rural counties and among for some racial/ethnic groups. More specifically, differential privacy could impact the ability to report on county-level death rates for the total population in rural areas and for subpopulations (e.g., by race and Hispanic ethnicity, age group) in many counties across the country.

This process would especially hamper surveillance of cardiovascular disease mortality within racial and Hispanic ethnic groups, many of which have higher mortality, and for younger adults, which have low but increasing cardiovascular disease mortality. Higher estimates in these smaller populations would mask the places and groups with truly high estimates. These potential problems with surveillance become magnified as these estimates are disseminated to state and local health departments for their program planning and resource allocation.

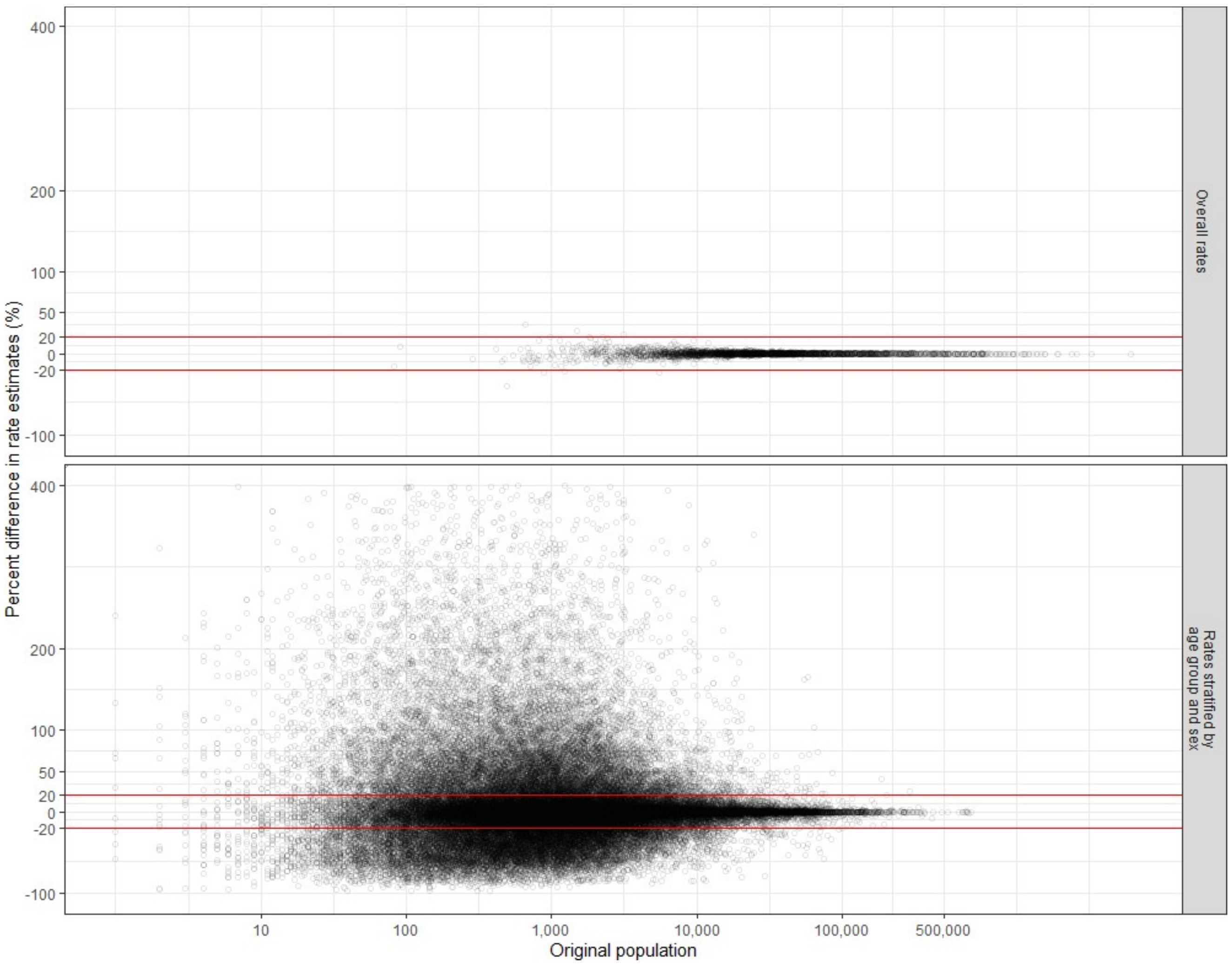


Figure 1: Percent difference in age-adjusted county-level heart disease death rates between NCHS populations and differential privacy population estimates. A positive difference indicates that the rate estimated using differential privacy populations was a higher value. Outliers have been truncated from this figure.

11. Susan Brower, Minnesota State Demographer

Dear Census Bureau,

Thank you for the opportunity to provide feedback on the 2010 demonstration data product for the Demographic and Housing Characteristics File (v. 2022-03-16).

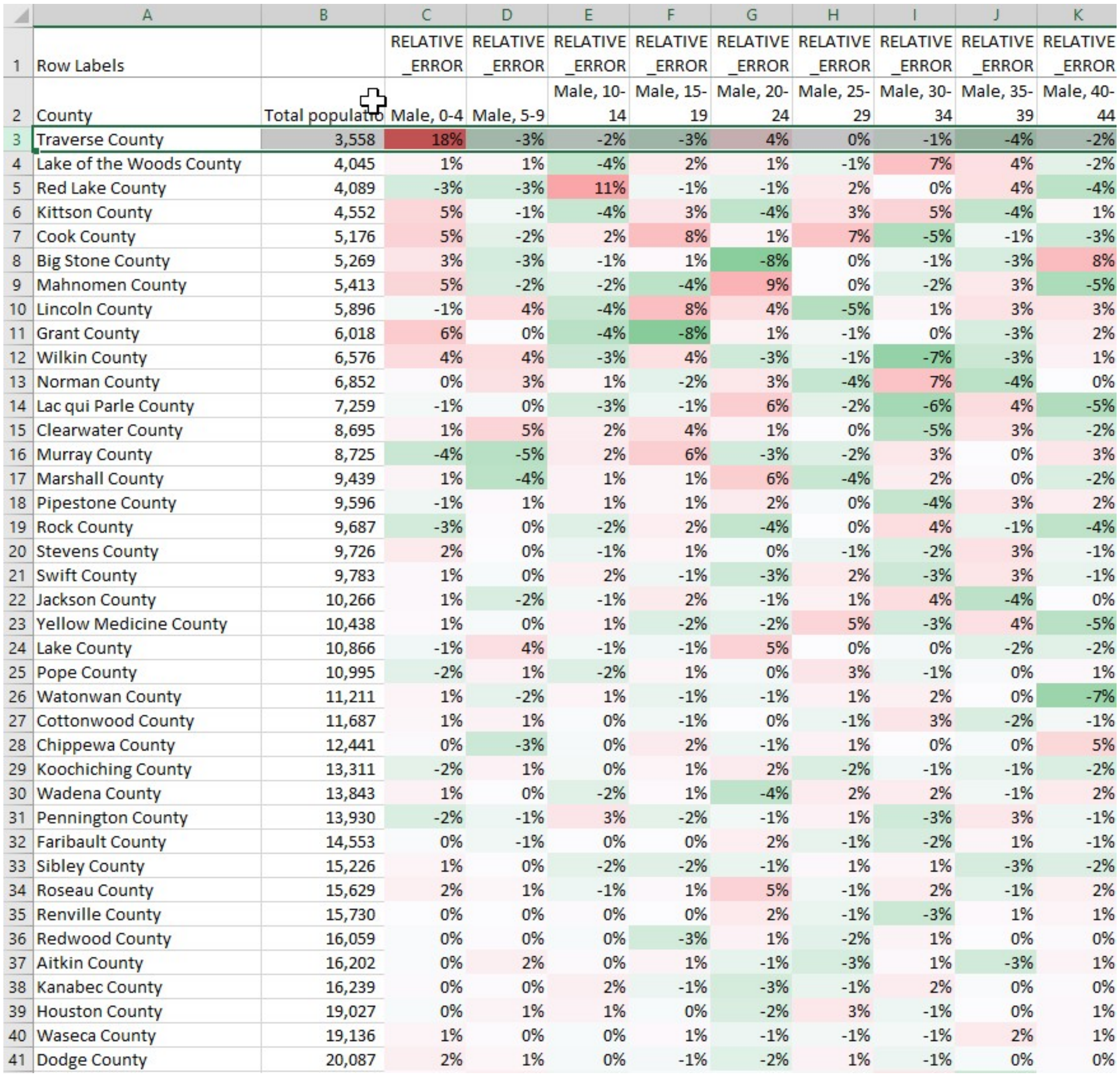
Our office is required by state statute to produce annual population projections at the county-level for each of Minnesota's 87 counties. The decennial census provides the base population for these projections. Decennial census data are also used as denominators in 5-year fertility and mortality rates which help form the foundation of our growth assumptions. In addition, we use decennial census data to create 5-year sex-specific progression ratios which are then used as additional inputs and checks. With these uses in mind, we focused our review of the demonstration data on 5-year age-sex groups for each of Minnesota's 87 counties.

In comparing age-sex data with the 2010 Summary Files and the 2010 differentially private demonstration data, we found:

- The majority of Minnesota's 87 counties (66%) had at least one age-sex group with a relative error of 3% or greater.
- In counties with a total population over 35,000, absolute relative errors greater than 3% were uncommon; however, the vast majority of counties in Minnesota (56 of 87 in 2020) have total populations below 35,000. Errors larger than 3% were common among these counties, with 88% of counties in this group with a relative error larger than 3% in at least one cell.
- In Minnesota's 10 smallest counties, the absolute relative error averaged between 3% and 4% for all cells in the table. The largest absolute relative errors in this group of counties were in the 18% to 22% range. While this level of error was uncommon, our projections require that age-sex data that be precise for all counties--including those will small populations. When age-sex data are used as progression ratios, large errors in one age group will corrupt the quality of older age groups as the model works its way forward. For this reason, we also need precise data for all cells in the age-sex table.

In conducting our review, we also found that errors tended to be greatest and most common in the four oldest age groups (ages 70-74, 75-79, 80-84 and 85+). These age groups were especially noisy when compared to younger age groups in all but the very largest counties. Precision in these older age groups is especially important not only for our projections modelling, but also for other state departments (Health, Pollution Control, etc.) which track and compare the prevalence and progression of cause-specific mortality. These departments also track racial and ethnic disparities in disease prevalence which I can only assume would be considerably noisier than the data for all race groups that I have looked at here.

I am attaching a worksheet showing the relative errors for the tables described above. Thank you for your consideration.

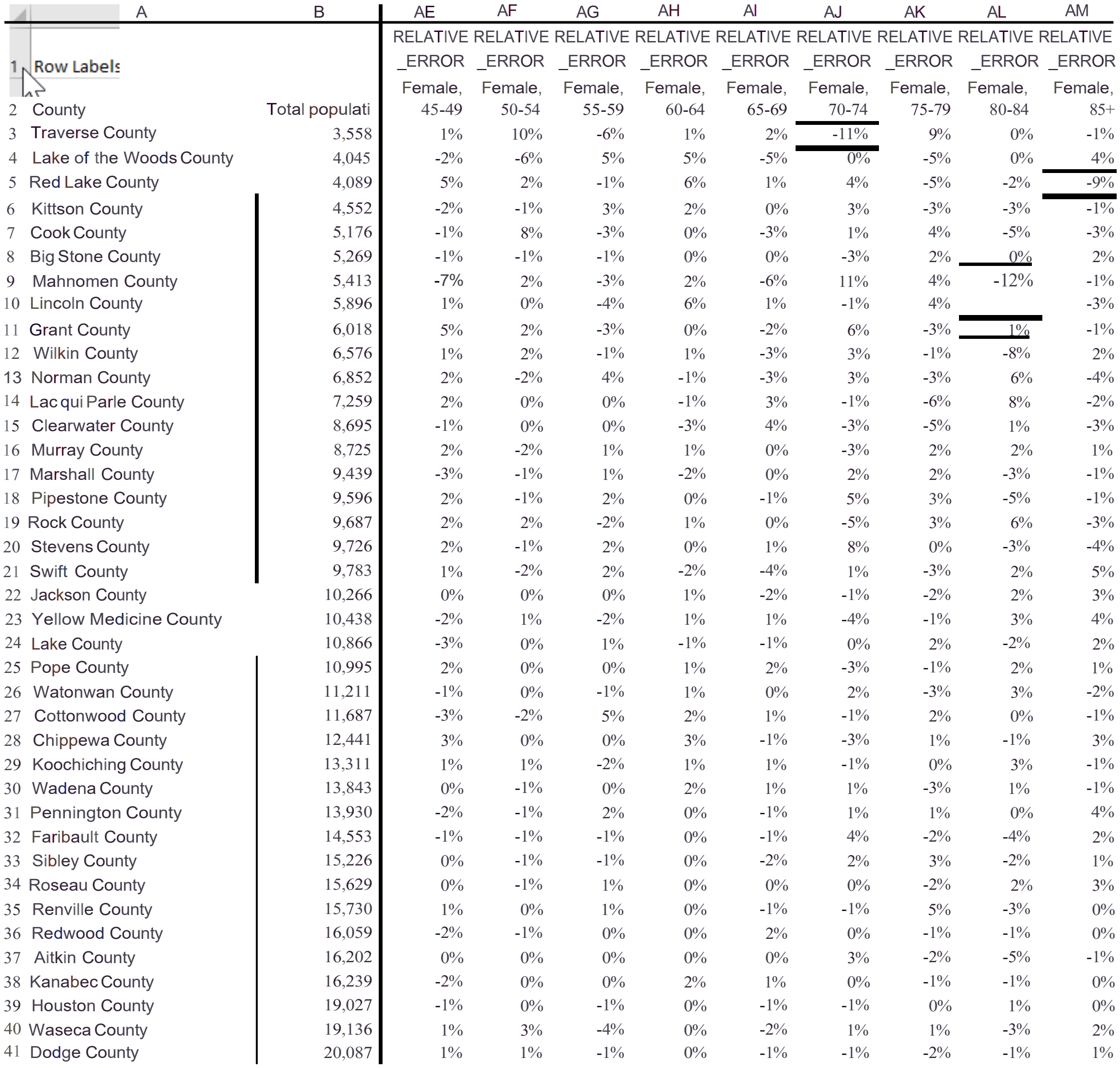


	A	B	C	D	E	F	G	H	I	J	K
1	Row Labels		RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR
2	County	Total populatio	Male, 0-4	Male, 5-9	Male, 10-14	Male, 15-19	Male, 20-24	Male, 25-29	Male, 30-34	Male, 35-39	Male, 40-44
3	Traverse County	3,558	18%	-3%	-2%	-3%	4%	0%	-1%	-4%	-2%
4	Lake of the Woods County	4,045	1%	1%	-4%	2%	1%	-1%	7%	4%	-2%
5	Red Lake County	4,089	-3%	-3%	11%	-1%	-1%	2%	0%	4%	-4%
6	Kittson County	4,552	5%	-1%	-4%	3%	-4%	3%	5%	-4%	1%
7	Cook County	5,176	5%	-2%	2%	8%	1%	7%	-5%	-1%	-3%
8	Big Stone County	5,269	3%	-3%	-1%	1%	-8%	0%	-1%	-3%	8%
9	Mahnomen County	5,413	5%	-2%	-2%	-4%	9%	0%	-2%	3%	-5%
10	Lincoln County	5,896	-1%	4%	-4%	8%	4%	-5%	1%	3%	3%
11	Grant County	6,018	6%	0%	-4%	-8%	1%	-1%	0%	-3%	2%
12	Wilkin County	6,576	4%	4%	-3%	4%	-3%	-1%	-7%	-3%	1%
13	Norman County	6,852	0%	3%	1%	-2%	3%	-4%	7%	-4%	0%
14	Lac qui Parle County	7,259	-1%	0%	-3%	-1%	6%	-2%	-6%	4%	-5%
15	Clearwater County	8,695	1%	5%	2%	4%	1%	0%	-5%	3%	-2%
16	Murray County	8,725	-4%	-5%	2%	6%	-3%	-2%	3%	0%	3%
17	Marshall County	9,439	1%	-4%	1%	1%	6%	-4%	2%	0%	-2%
18	Pipestone County	9,596	-1%	1%	1%	1%	2%	0%	-4%	3%	2%
19	Rock County	9,687	-3%	0%	-2%	2%	-4%	0%	4%	-1%	-4%
20	Stevens County	9,726	2%	0%	-1%	1%	0%	-1%	-2%	3%	-1%
21	Swift County	9,783	1%	0%	2%	-1%	-3%	2%	-3%	3%	-1%
22	Jackson County	10,266	1%	-2%	-1%	2%	-1%	1%	4%	-4%	0%
23	Yellow Medicine County	10,438	1%	0%	1%	-2%	-2%	5%	-3%	4%	-5%
24	Lake County	10,866	-1%	4%	-1%	-1%	5%	0%	0%	-2%	-2%
25	Pope County	10,995	-2%	1%	-2%	1%	0%	3%	-1%	0%	1%
26	Watonwan County	11,211	1%	-2%	1%	-1%	-1%	1%	2%	0%	-7%
27	Cottonwood County	11,687	1%	1%	0%	-1%	0%	-1%	3%	-2%	-1%
28	Chippewa County	12,441	0%	-3%	0%	2%	-1%	1%	0%	0%	5%
29	Koochiching County	13,311	-2%	1%	0%	1%	2%	-2%	-1%	-1%	-2%
30	Wadena County	13,843	1%	0%	-2%	1%	-4%	2%	2%	-1%	2%
31	Pennington County	13,930	-2%	-1%	3%	-2%	-1%	1%	-3%	3%	-1%
32	Faribault County	14,553	0%	-1%	0%	0%	2%	-1%	-2%	1%	-1%
33	Sibley County	15,226	1%	0%	-2%	-2%	-1%	1%	1%	-3%	-2%
34	Roseau County	15,629	2%	1%	-1%	1%	5%	-1%	2%	-1%	2%
35	Renville County	15,730	0%	0%	0%	0%	2%	-1%	-3%	1%	1%
36	Redwood County	16,059	0%	0%	0%	-3%	1%	-2%	1%	0%	0%
37	Aitkin County	16,202	0%	2%	0%	1%	-1%	-3%	1%	-3%	1%
38	Kanabec County	16,239	0%	0%	2%	-1%	-3%	-1%	2%	0%	0%
39	Houston County	19,027	0%	1%	1%	0%	-2%	3%	-1%	0%	1%
40	Waseca County	19,136	1%	0%	0%	1%	-1%	-1%	-1%	2%	1%
41	Dodge County	20,087	2%	1%	0%	-1%	-2%	1%	-1%	0%	0%

	A	8	M	N	0	P	Q	R	S	T
	Row Labels 0		RELATIVE _ERROR Male,45-	RELATIVE _ERROR Male, SO-	RELATIVE _ERROR Male, 55-	RELATIVE _ERROR Male, 60-	RELATIVE _ERROR Male, 65-	RELATIVE _ERROR Male, 70-	RELATIVE _ERROR Male, 75-	RELATIVE _ERROR Male, 80-	RELATIVE _ERROR
2	County	Total populati	49	54	59	64	69	74	79	84	~~Male, 85+~~
3	Traverse County	3,558	-1%	4%	-4%	-1%	-2%	1%	4%	-2%	-3%
4	Lake of the Woods County	4,045	-1%	-3%	2%	-1%	3%	-3%	6%	0%	-11%
5	Red Lake County	4,089	2%	0%	2%	-5%	-4%	0%	-4%	2%	18%
6	Kittson County	4,552	-3%	-1%	1%	2%	-3%	7%		0%	6%
7	Cook County	5,176	3%	-2%	1%	-1%	-6%	-6%	3%	7%	-4%
8	Big Stone County	5,269	-5%	2%	1%	0%	2%	2%	0%	-3%	-5%
9	Mahnomen County	5,413	0%	-1%	2%	-1%	4%			-8%	3%
10	Lincoln County	5,896	-4%	1%	-3%	1%	-1%			6%	-3%
11	Grant County	6,018	1%	4%	-4%	3%	1%			8%	1%
12	Wilkin County	6,576	-1%	1%	1%	-1%	6%	0%	0%	-4%	-4%
13	Norman County	6,852	0%	-3%	0%	3%	-2%	4%	0%	0%	3%
14	Lac qui Parle County	7,259	0%	1%	3%	1%	0%	0%	1%	1%	-1%
15	Clearwater County	8,695	-3%	0%	2%	1%	-1%	-5%	-2%	2%	7%
16	Murray County	8,725	-1%	1%	-2%	-2%	2%	-1%	-1%	7%	-1%
17	Marshall County	9,439	1%	1%	-1%	0%	-3%	-1%	-1%	-2%	8%
18	Pipestone County	9,596	-1%	-3%	2%	0%	1%	-5%	1%	6%	-5%
19	Rock County	9,687	4%	-1%	2%	-1%	-1%	4%	2%1	-7%	7%
20	Stevens County	9,726	1%	2%	0%	2%	0%	2%	-2%	-§2.f	1%
21	Swift County	9,783	2%	0%	0%	-2%	-1%	-2%	3%	0%	2%
22	Jackson County	10,266	0%	-3%	1%	0%	0%	2%	-2%	1%	-1%
23	Yellow Medicine County	10,438	2%	1%	-2%	2%	1%	9%	-1%[	-1	-1%
24	Lake County	10,866	2%	-1%	-1%	2%	2%	-2%	-3%	4%	-1%
25	Pope County	10,995	-2%	0%	1%	-1%	1%	1%	1%	0%	-1%
26	Watonwan County	11,211	4%	1%	0%	1%	-4%	-2%	0%	3%	5%
27·	Cottonwood County	11,687	-1%	1%	0%	-1%	-1%	-1%	1%	-1%	2%
28	Chippewa County	12,441	-4%	-2%	2%	-1%	1%	-1%	4%	-5%	1%
29	Koochiching County	13,311	3%	1%	-3%	2%	-3%	0%	1%	1%	0%
30	Wadena County	13,843	-2%	1%	0%	-1%	1%	-1%	1%	0%	2%
31	Pennington County	13,930	1%	0%	0%	1%	2%	-4%	-1%	13%	-5%
32	Faribault County	14,553	2%	0%	1%	0%	1%	-1%	0%	3%	0%
33	Sibley County	15,226	2%	0%	1%	1%	1%	-2%	2%	-1%	6%
34	Roseau County	15,629	0%	-1%	-1%	-1%	-1%	0%	-2%	3%	-5%
35	Renville County	15,730	0%	-2%	1%	2%	-1%	1%	-4%	-1%	4%
36	Redwood County	16,059	-2%	1%	1%	0%	1%	-1%	2%	-1%	5%
37	Aitkin County	16,202	-1%	1%	0%	0%	1%	0%	-1%	3%	-5%
38	Kanabec County	16,239	1%	-2%	1%	1%	1%	-4%	-2%	3%	0%
39	Houston County	19,027	0%	0%	-1%	1%	0%	0%	1%	1%	-4%
40	Waseca County	19,136	0%	0%	-2%	1%	-1%	1%	-3%	1%	-1%
41	Dodge County	20,087	0%	0%	0%	0%	0%	0%	-4%	5%	-2%

	A	B	V	W	X	y	Z	AA	AB	AC	AD
			RELATIVE	RELATIVE	RELATIVE	RELATIVE	RELATIVE	RELATIVE	RELATIVE	RELATIVE	RELATIVE
	Row Labels		_ERROR	_ERROR	_ERROR	_ERROR	_ERROR	_ERROR	_ERROR	_ERROR	_ERROR
			Female,	Female,	Female,	Female,	Female,	Female,	Female,	Female,	Female,
2	County	Total populati	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44
3	Traverse County	3,558	-3%	3%	3%	-11%	4%	0%	-3%	2%	0%
4	Lake of the Woods County	4,045	1%	4%	-3%	1%	3%	-5%	-8%	3%	4%
5	Red Lake County	4,089	-3%	-1%	1%	-1%	3%	-2%	-2%	4%	-8%
6	Kittson County	4,552	-3%	-1%	-2%	3%	0%	0%	1%	0%	-3%
7	Cook County	5,176	-2%	4%	5%	0%	-6%	5%	-2%	-4%	-1%
8	Big Stone County	5,269	-1%	0%	3%	-3%j	ml	-!!	-2%	6%	-1%
9	Mahnomen County	5,413	-4%	4%	-2%	9%	1%	0%	1%	-5%	6%
10	Lincoln County	5,896	3%	-1%	-2%	7%	0%	2%	3%i	-!!:/§.	-2%
11	Grant County	6,018	-1%	1%	-1%	2%	1%	1%	-4%	0%	-5%
12	Wilkin County	6,576	0%	3%	3%	0%	5%	-4%	1%	-7%	1%
13	Norman County	6,852	1%	-2%	3%	-4%	8%j	-12%1	1%	2%	1%
14	Lac qui Parle County	7,259	1%	1%	0%	0%	1%	-1%	-2%	2%	-3%
15	Clearwater County	8,695	-1%	-3%	-2%	5%	5%	-2%	-3%	0%	0%
16	Murray County	8,725	-1%	-1%	0%	1%	1%	3%	0%	-5%	-3%
17	Marshall County	9,439	-2%	3%	-2%	2%	-3%	1%	4%	2%	2%
18	Pipestone County	9,596	1%	-4%	-2%	0%	1%	0%	-4%	2%	-1%
19	Rock County	9,687	-1%	-3%	1%	3%	-3%	-1%	2%	-1%	0%
20	Stevens County	9,726	-2%	2%	1%	-2%	0%	-2%	-2%	2%	-1%
21	Swift County	9,783	-4%	2%	1%	0%	-1%	-2%	3%	1%	0%
22	Jackson County	10,266	0%	1%	2%	-4%	2%	-3%	7%	-3%	1%
23	Yellow Medicine County	10,438	1%	2%	-5%	3%	-5%	2%	1%	-3%	1%
24	Lake County	10,866	1%	-1%	1%	-3%	1%	-2%	5%	-3%	-1%
25	Pope County	10,995	2%	-3%	1%	0%	-7%	2%	1%	-3%	-3%
26	Watonwan County	11,211	2%	-2%	-3%	2%	0%	1%	3%	-2%	0%
27	Cottonwood County	11,687	0%	-1%	0%	-2%	-8%	6%	-1%	2%	2%
28	Chippewa County	12,441	-1%	1%	0%	1%	2%	-4%	-4%	5%	-4%
29	Koochiching County	13,311	2%	-1%	2%	-3%	5%	-2%	2%	-2%	1%
30	Wadena County	13,843	-1%	1%	-2%	1%	-3%	5%	-1%	-1%	1%
31	Pennington County	13,930	-2%	0%	-4%	1%	-2%	1%	-1%	2%	1%
32	Faribault County	14,553	3%	-2%	-3%	0%	4%	-1%	-2%	2%	0%
33	Sibley County	15,226	-2%	-2%	1%	2%	0%	2%	2%	-2%	2%
34	Roseau County	15,629	0%	-1%	1%	-3%	-1%	0%	2%	-2%	1%
35	Renville County	15,730	3%	1%	-1%	0%	1%	-2%	-1%	0%	-3%
36	Redwood County	16,059	-1%	-2%	2%	3%	-2%	1%	4%	-2%	-1%
37	Aitkin County	16,202	0%	-2%	1%	-1%	-2%	2%	1%	0%	1%
38	Kanabec County	16,239	-2%	1%	1%	0%	-2%	-1%	0%	0%	3%
39	Houston County	19,027	2%	1%	-1%	-1%	-1%	0%	0%	0%	1%
40	Waseca County	19,136	0%	0%	-1%	3%	-3%	0%	1%	0%	-1%
41	Dodge County	20,087									

0%	0%	0%	-1%
	0%	1%	-1%
	0%	-1%	

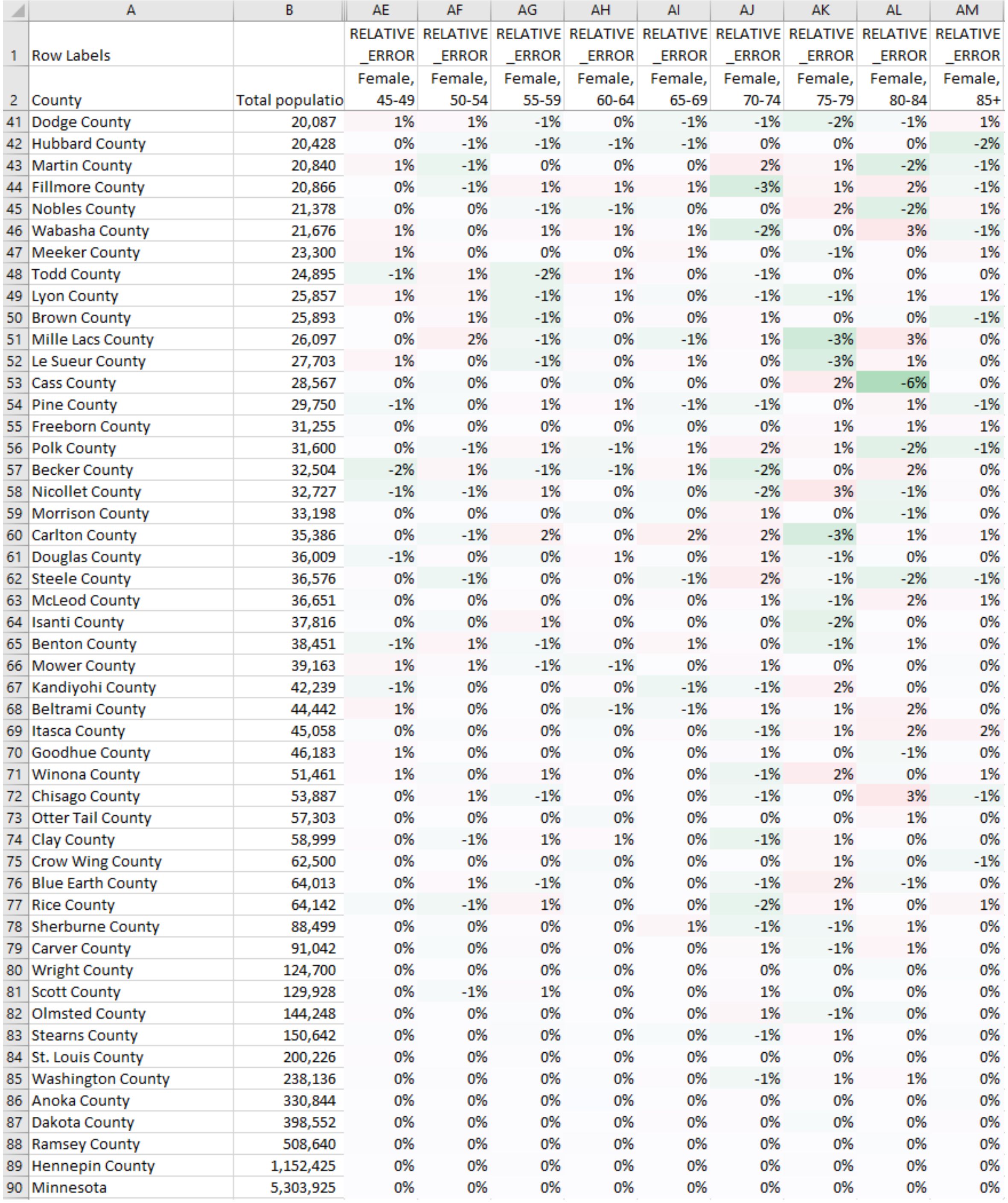


	A	B	AE	AF	AG	AH	AI	AJ	AK	AL	AM
1	Row Labels		RELATIVE _ERROR Female,	RELATIVE _ERROR Female,	RELATIVE _ERROR Female,	RELATIVE _ERROR Female,	RELATIVE _ERROR Female,	RELATIVE _ERROR Female,	RELATIVE _ERROR Female,	RELATIVE _ERROR Female,	RELATIVE _ERROR Female,
2	County	Total populati	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85+
3	Traverse County	3,558	1%	10%	-6%	1%	2%	-11%	9%	0%	-1%
4	Lake of the Woods County	4,045	-2%	-6%	5%	5%	-5%	0%	-5%	0%	4%
5	Red Lake County	4,089	5%	2%	-1%	6%	1%	4%	-5%	-2%	-9%
6	Kittson County	4,552	-2%	-1%	3%	2%	0%	3%	-3%	-3%	-1%
7	Cook County	5,176	-1%	8%	-3%	0%	-3%	1%	4%	-5%	-3%
8	Big Stone County	5,269	-1%	-1%	-1%	0%	0%	-3%	2%	0%	2%
9	Mahnomen County	5,413	-7%	2%	-3%	2%	-6%	11%	4%	-12%	-1%
10	Lincoln County	5,896	1%	0%	-4%	6%	1%	-1%	4%		-3%
11	Grant County	6,018	5%	2%	-3%	0%	-2%	6%	-3%	1%	-1%
12	Wilkin County	6,576	1%	2%	-1%	1%	-3%	3%	-1%	-8%	2%
13	Norman County	6,852	2%	-2%	4%	-1%	-3%	3%	-3%	6%	-4%
14	Lac qui Parle County	7,259	2%	0%	0%	-1%	3%	-1%	-6%	8%	-2%
15	Clearwater County	8,695	-1%	0%	0%	-3%	4%	-3%	-5%	1%	-3%
16	Murray County	8,725	2%	-2%	1%	1%	0%	-3%	2%	2%	1%
17	Marshall County	9,439	-3%	-1%	1%	-2%	0%	2%	2%	-3%	-1%
18	Pipestone County	9,596	2%	-1%	2%	0%	-1%	5%	3%	-5%	-1%
19	Rock County	9,687	2%	2%	-2%	1%	0%	-5%	3%	6%	-3%
20	Stevens County	9,726	2%	-1%	2%	0%	1%	8%	0%	-3%	-4%
21	Swift County	9,783	1%	-2%	2%	-2%	-4%	1%	-3%	2%	5%
22	Jackson County	10,266	0%	0%	0%	1%	-2%	-1%	-2%	2%	3%
23	Yellow Medicine County	10,438	-2%	1%	-2%	1%	1%	-4%	-1%	3%	4%
24	Lake County	10,866	-3%	0%	1%	-1%	-1%	0%	2%	-2%	2%
25	Pope County	10,995	2%	0%	0%	1%	2%	-3%	-1%	2%	1%
26	Watonwan County	11,211	-1%	0%	-1%	1%	0%	2%	-3%	3%	-2%
27	Cottonwood County	11,687	-3%	-2%	5%	2%	1%	-1%	2%	0%	-1%
28	Chippewa County	12,441	3%	0%	0%	3%	-1%	-3%	1%	-1%	3%
29	Koochiching County	13,311	1%	1%	-2%	1%	1%	-1%	0%	3%	-1%
30	Wadena County	13,843	0%	-1%	0%	2%	1%	1%	-3%	1%	-1%
31	Pennington County	13,930	-2%	-1%	2%	0%	-1%	1%	1%	0%	4%
32	Faribault County	14,553	-1%	-1%	-1%	0%	-1%	4%	-2%	-4%	2%
33	Sibley County	15,226	0%	-1%	-1%	0%	-2%	2%	3%	-2%	1%
34	Roseau County	15,629	0%	-1%	1%	0%	0%	0%	-2%	2%	3%
35	Renville County	15,730	1%	0%	1%	0%	-1%	-1%	5%	-3%	0%
36	Redwood County	16,059	-2%	-1%	0%	0%	2%	0%	-1%	-1%	0%
37	Aitkin County	16,202	0%	0%	0%	0%	0%	3%	-2%	-5%	-1%
38	Kanabec County	16,239	-2%	0%	0%	2%	1%	0%	-1%	-1%	0%
39	Houston County	19,027	-1%	0%	-1%	0%	-1%	-1%	0%	1%	0%
40	Waseca County	19,136	1%	3%	-4%	0%	-2%	1%	1%	-3%	2%
41	Dodge County	20,087	1%	1%	-1%	0%	-1%	-1%	-2%	-1%	1%

	A		C	D			G	H			K
	Row Labels		RELATIVE	RELATIVE	RELATIVE	RELATIVE	RELATIVE	RELATIVE	RELATIVE	RELATIVE	RELATIVE
			_ERROR	_ERROR	_ERROR	_ERROR	_ERROR	_ERROR	_ERROR	_ERROR	_ERROR
					Male, 10-	Male, 15-	Male, 20-	Male, 25-	Male, 30-	Male, 35-	Male,40-
2	County	Total populatio	Male, 0-4	Male,5-9	M	e	M		M		44
41	Dodge County	20,087	2%	1%	0%	**-1%**	-2%	**1%**	-1%	0%	0%
42	Hubbard County	20,428	1%	2%	**-1%**	0%	0%	**-1%**	3%	-3%	0%
43	Martin County	20,840	0%	0%	0%	**-1%**	2%	**-4%**	-1%	2%	0%
44	Fillmore County	20,866	-1%	1%	0%	1%	2%	-3%	1%	-2%	0%
45	Nobles County	21,378	0%	1%	-2%	0%	2%	**-1%**	0%	0%	0%
46	Wabasha County	21,676	1%	-1%	0%	**-1%**	0%	1%	-1%	**-1%**	0%
47	Meeker County	23,300	-1%	0%	**1%**	**-1%**	3%	-3%	-1%	0%	1%
48	Todd County	24,895	1%	1%	0%	0%	-5%	**4%**	-2%	3%	-1%
49	Lyon County	25,857	-1%	1%	**1%**	**-1%**	-1%	2%	1%	0%	0%
SO	Brown County	25,893	0%	0%	0%	**1%**	-1%	**1%**	1%	0%	-1%
51	Mille Lacs County	26,097	1%	-1%	**-1%**	-2%	-3%	**1%**	0%	**1%**	1%
52	Le Sueur County	27,703	-1%	-1%	**1%**	0%	2%	-2%	1%	-2%	-1%
53	Cass County	28,567	0%	1%	**-1%**	-2%	-3%	**1%**	2%	0%	1%
54	Pine County	29,750	0%	1%	0%	-2%	-3%	0%	1%	**-1%**	1%
55	Freeborn County	31,255	0%	-1%	**1%**	**1%**	1%	**-1%**	0%	**-1%**	0%
56	Polk County	31,600	0%	0%	**1%**	**-1%**	-1%	**-1%**	-3%	1%	1%
57	Becker County	32,504	-1%	0%	**1%**	**-1%**	0%	0%	-1%	**1%**	1%
58	Nicollet County	32,727	0%	-1%	0%	-2%	1%	**-1%**	0%	0%	1%
59	Morrison County	33,198	1%	0%	**-1%**	0%	1%	**-1%**	0%	0%	0%
60	Carlton County	35,386	-1%	0%	0%	-2%	0%	0%	-2%	2%	-2%
61	Douglas County	36,009	0%	2%	0%	**-1%**	-1%	0%	0%	**1%**	0%
62	Steele County	36,576	0%	0%	0%	**-1%**	1%	0%	-1%	0%	2%
63	Mcleod County	36,651	-1%	0%	0%	0%	-1%	**1%**	-2%	**1%**	1%
64	Isanti County	37,816	-1%	0%	**1%**	**1%**	0%	**1%**	-1%	2%	0%
65	Benton County	38,451	-1%	-1%	0%	0%	1%	**-1%**	0%	0%	1%
66	Mower County	39,163	0%	-1%	**1%**	**-1%**	1%	0%	0%	**1%**	0%
67	Kandiyohi County	42,239	1%	1%	0%	**-1%**	-2%	2%	1%	0%	0%
68	Beltrami County	44,442	0%	1%	0%	0%	-1%	0%	2%	-2%	0%
69	Itasca County	45,058	0%	-1%	0%	-2%	-2%	0%	1%	**-1%**	1%
70	Goodhue County	46,183	0%	0%	**-1%**	**1%**	0%	**-1%**	0%	0%	0%
71	Winona County	51,461	0%	-1%	**1%**	**-1%**	-1%	2%	0%	**1%**	0%
72	Chisago County	53,887	0%	0%	0%	0%	0%	0%	1%	0%	0%
73	Otter Tail County	57,303	0%	-1%	0%	0%	-1%	0%	1%	0%	0%
74	Clay County	58,999	0%	0%	0%	**-1%**	-1%	**1%**	1%	0%	0%
75	Crow Wing County	62,500	0%	-1%	0%	0%	-1%	**1%**	0%	0%	0%
76	Blue Earth County	64,013	0%	0%	**1%**	**-1%**	-1%	**1%**	0%	0%	-1%
77	Rice County	64,142	0%	0%	**-1%**	0%	-1%	2%	-1%	**1%**	1%
78	Sherburne County	88,499	0%	0%	0%	0%	-1%	0%	0%	0%	-1%
79	Carver County	91,042	0%	0%	0%	0%	1%	0%	-1%	0%	0%
80	Wright County	124,700	0%	0%	0%	0%	0%	0%	0%	0%	0%
81	Scott County	129,928	0%	0%	0%	0%	0%	0%	0%	0%	0%
82	Olmsted County	144,248	0%	0%	0%	**-1%**	-1%	0%	0%	**1%**	0%
83	Stearns County	150,642	0%	0%	0%	0%	0%	0%	0%	0%	0%
84	St. Louis County	200,226	0%	0%	0%	**-1%**	0%	0%	0%	0%	0%
85	Washington County	238,136	0%	0%	0%	0%	-1%	0%	0%	0%	0%
86	Anoka County	330,844	0%	0%	0%	0%	0%	0%	0%	0%	0%
87	Dakota County	398,552	0%	0%	0%	0%	0%	0%	0%	0%	0%
88	Ramsey County	508,640	0%	0%	0%	**-1%**	0%	0%	0%	0%	0%
89	Hennepin County	1,152,425	0%	0%	0%	0%	0%	0%	0%	0%	0%
90	Minnesota	5,303,925	0%	0%	0%	0%	0%	0%	0%	0%	0%

	A			M	N	0		Q	R	S	I
	Row Labels		RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR
2	County	Total populatio	Male, 45- 49	Male, SO- 54	Male, 55- 59	Male, 60- 64	Male, 65- 69	Male, 70- 74	Male, 75- 79	Male, 80- 84	Male, 85+
41	Dodge County	20,087	0%	0%	0%	0%	0%	0%	-4%	5%	-2%
42	Hubbard County	20,428	0%	2%	-1%	-1%	1%	-2%	-1%	3%	4%
43	Martin County	20,840	1%	0%	-1%	-1%	1%	4%	-1%	-2%	1%
44	Fillmore County	20,866	0%	0%	0%	0%	-1%	1%	0%	0%	2%
45	Nobles County	21,378	0%	1%	-1%	0%	2%	-1%	-1%	3%	0%
46	Wabasha County	21,676	0%	0%	-1%	1%	1%	-1%	-2%	3%	2%
47	Meeker County	23,300	-1%	-1%	1%	0%	0%	3%	2%	-4%	0%
48	Todd County	24,895	2%	0%	1%	0%	0%	3%	-2%	1%	-4%
49	Lyon County	25,857	-1%	1%	0%	0%	0%	1%	2%	0%	-1%
SO	Brown County	25,893	1%	0%	-1%	0%	0%	-1%	1%	-1%	1%
51	Mille Lacs County	26,097	1%	1%	-1%	-1%	-1%	-5%	2%	5%	3%
52	Le Sueur County	27,703	1%	0%	0%	1%	0%	1%	-1%	-1%	1%
53	Cass County	28,567	-1%	-1%	1%	1%	0%	0%	1%	-3%	0%
54	Pine County	29,750	-1%	0%	-1%	1%	1%	0%	0%	-2%	-1%
55	Freeborn County	31,255	1%	1%	-1%	-1%	1%	-3%	0%	3%	0%
56	Polk County	31,600	2%	0%	1%	0%	0%	0%	1%	0%	-2%
57	Becker County	32,504	-1%	-1%	0%	0%	1%	0%	0%	0%	0%
58	Nicollet County	32,727	-1%	-1%	1%	-1%	0%	0%	0%	0%	3%
59	Morrison County	33,198	-1%	0%	0%	1%	0%	-2%	3%	3%	-3%
60	Carlton County	35,386	2%	0%	1%	1%	1%	3%	-1%	-2%	-4%
61	Douglas County	36,009	0%	0%	0%	1%	1%	-1%	-1%	-1%	-1%
62	Steele County	36,576	-1%	0%	-1%	1%	0%	-1%	1%	0%	1%
63	Mcleod County	36,651	0%	1%	0%	0%	0%	2%	-2%	-1%	1%
64	Isanti County	37,816	0%	0%	0%	0%	-1%	-1%	0%	2%	2%
65	Benton County	38,451	-1%	1%	0%	1%	0%	0%	0%	2%	-3%
66	Mower County	39,163	-1%	0%	1%	-1%	0%	2%	-1%	0%	1%
67	Kandiyohi County	42,239	0%	0%	1%	0%	0%	2%	0%	0%	0%
68	Beltrami County	44,442	0%	0%	0%	0%	1%	0%	-1%	6%	-5%
69	Itasca County	45,058	-1%	1%	0%	0%	1%	1%	0%	-1%	1%
70	Goodhue County	46,183	0%	0%	0%	1%	0%	-1%	-2%	2%	2%
71	Winona County	51,461	0%	1%	0%	0%	0%	1%	3%	-2%	-2%
72	Chisago County	53,887	0%	0%	0%	-1%	0%	1%	-1%	-1%	2%
73	Otter Tail County	57,303	0%	1%	-1%	0%	0%	-1%	2%	-2%	1%
74	Clay County	58,999	1%	0%	0%	0%	0%	-1%	1%	2%	0%
75	Crow Wing County	62,500	0%	0%	0%	0%	0%	0%	1%	1%	-1%
76	Blue Earth County	64,013	1%	0%	1%	0%	0%	0%	1%	1%	-1%
77	Rice County	64,142	-1%	0%	0%	0%	0%	0%	0%	1%	1%
78	Sherburne County	88,499	1%	0%	1%	0%	1%	0%	0%	-3%	-1%
79	Carver County	91,042	0%	0%	0%	0%	0%	0%	1%	-1%	1%
80	Wright County	124,700	0%	0%	0%	0%	1%	-1%	0%	1%	0%
81	Scott County	129,928	0%	0%	0%	0%	0%	1%	1%	-5%	-1%
82	Olmsted County	144,248	0%	0%	0%	0%	0%	0%	-1%	1%	0%
83	Stearns County	150,642	0%	0%	0%	0%	0%	0%	1%	-1%	0%
84	St. Louis County	200,226	0%	0%	0%	0%	0%	0%	0%	0%	0%
85	Washington County	238,136	0%	0%	0%	0%	0%	1%	-1%	0%	0%
86	Anoka County	330,844	0%	0%	0%	0%	0%	0%	0%	-1%	-1%
87	Dakota County	398,552	0%	0%	0%	0%	0%	0%	0%	0%	0%
88	Ramsey County	508,640	0%	0%	0%	0%	0%	1%	0%	0%	0%
89	Hennepin County	1,152,425	0%	0%	0%	0%	0%	0%	0%	0%	0%
90	Minnesota	5,303,925	0%	0%	0%	0%	0%	0%	0%	0%	0%

	A		V	W	X	y	Z	AA	AB	AC	AD
	Row Labels										
2	County	Total populatio	RELATIVE _ERROR Female, 0-4	RELATIVE _ERROR Female, 5-9	RELATIVE _ERROR Female, 10-14	RELATIVE _ERROR Female, 15-19	RELATIVE _ERROR Female, 20-24	RELATIVE _ERROR Female, 25-29	RELATIVE _ERROR Female, 30-34	RELATIVE _ERROR Female, 35-39	RELATIVE _ERROR Female, 40-44
41	Dodge County	20,087	0%	0%	0%	-1%	0%	1%	-1%	0%	-1%
42	Hubbard County	20,428	1%	1%	0%	2%	6%	-4%	3%	-2%	-1%
43	Martin County	20,840	0%	-1%	0%	0%	-1%	2%	0%	0%	-1%
44	Fillmore County	20,866	-1%	0%	-1%	2%	-2%	1%	0%	-1%	-1%
45	Nobles County	21,378	-1%	1%	-1%	0%	-2%	0%	1%	-2%	0%
46	Wabasha County	21,676	-2%	0%	0%	1%	2%	-3%	1%	-1%	-2%
47	Meeker County	23,300	1%	0%	-1%	0%	0%	-1%	-1%	0%	-1%
48	Todd County	24,895	0%	0%	-1%	0%	-1%	-1%	-1%	0%	2%
49	Lyon County	25,857	0%	1%	-1%	0%	-1%	1%	0%	0%	-1%
SO	Brown County	25,893	0%	1%	1%	0%	0%	0%	0%	-1%	0%
51	Mille Lacs County	26,097	0%	0%	1%	2%	0%	-1%	1%	-2%	1%
52	Le Sueur County	27,703	1%	0%	0%	0%	0%	-1%	-3%	2%	-1%
53	Cass County	28,567	1%	1%	-1%	0%	2%	-1%	0%	1%	0%
54	Pine County	29,750	2%	0%	1%	0%	-2%	1%	1%	0%	0%
55	Freeborn County	31,255	-1%	1%	-1%	-2%	0%	-1%	2%	-2%	1%
56	Polk County	31,600	0%	-1%	1%	1%	1%	-3%	0%	0%	0%
57	Becker County	32,504	0%	0%	0%	0%	1%	0%	0%	1%	2%
58	Nicollet County	32,727	-1%	1%	0%	0%	1%	1%	0%	1%	1%
59	Morrison County	33,198	0%	-1%	0%	-1%	-1%	1%	-1%	0%	1%
60	Carlton County	35,386	0%	-1%	0%	0%	3%	-2%	0%	0%	1%
61	Douglas County	36,009	-1%	0%	1%	0%	1%	-1%	1%	-1%	0%
62	Steele County	36,576	1%	0%	0%	1%	-1%	1%	1%	0%	0%
63	Mcleod County	36,651	0%	-1%	1%	-1%	0%	0%	0%	1%	0%
64	Isanti County	37,816	0%	-1%	0%	-1%	1%	0%	0%	0%	1%
65	Benton County	38,451	0%	0%	0%	-1%	-1%	2%	-1%	1%	1%
66	Mower County	39,163	1%	-1%	0%	-1%	1%	-1%	-1%	1%	0%
67	Kandiyohi County	42,239	0%	0%	0%	0%	0%	-1%	0%	0%	1%
68	Beltrami County	44,442	1%	-1%	0%	-1%	0%	0%	0%	1%	0%
69	Itasca County	45,058	0%	1%	1%	-1%	-1%	1%	1%	0%	0%
70	Goodhue County	46,183	0%	0%	1%	-1%	-1%	1%	-1%	1%	0%
71	Winona County	51,461	-1%	0%	0%	0%	0%	0%	0%	-1%	0%
72	Chisago County	53,887	0%	0%	0%	-1%	1%	0%	1%	0%	0%
73	Otter Tail County	57,303	-1%	0%	1%	0%	1%	-1%	-2%	2%	-1%
74	Clay County	58,999	0%	0%	1%	0%	-1%	1%	-1%	0%	0%
75	Crow Wing County	62,500	0%	0%	0%	0%	0%	1%	0%	-1%	0%
76	Blue Earth County	64,013	0%	0%	0%	0%	0%	1%	0%	1%	0%
77	Rice County	64,142	0%	0%	0%	0%	0%	0%	0%	0%	1%
78	Sherburne County	88,499	0%	0%	0%	0%	-1%	0%	0%	0%	0%
79	Carver County	91,042	0%	0%	0%	0%	1%	-1%	0%	0%	0%
80	Wright County	124,700	0%	0%	0%	0%	0%	0%	0%	0%	0%
81	Scott County	129,928	0%	0%	0%	0%	-2%	1%	0%	0%	0%
82	Olmsted County	144,248	0%	0%	0%	-1%	0%	0%	0%	0%	0%
83	Stearns County	150,642	0%	0%	0%	0%	0%	0%	-1%	0%	0%
84	St. Louis County	200,226	0%	0%	0%	0%	0%	0%	0%	0%	0%
85	Washington County	238,136	0%	0%	0%	0%	-1%	1%	0%	0%	0%
86	Anoka County	330,844	0%	0%	0%	0%	0%	0%	0%	0%	0%
87	Dakota County	398,552	0%	0%	0%	0%	-1%	0%	0%	0%	0%
88	Ramsey County	508,640	0%	0%	0%	0%	0%	0%	0%	0%	0%
89	Hennepin County	1,152,425	0%	0%	0%	0%	0%	0%	0%	0%	0%
90	Minnesota	5,303,925	0%	0%	0%	0%	0%	0%	0%	0%	0%



	A	B	AE	AF	AG	AH	AI	AJ	AK	AL	AM
1	Row Labels		RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR	RELATIVE _ERROR
2	County	Total populatio	Female, 45-49	Female, 50-54	Female, 55-59	Female, 60-64	Female, 65-69	Female, 70-74	Female, 75-79	Female, 80-84	Female, 85+
41	Dodge County	20,087	1%	1%	-1%	0%	-1%	-1%	-2%	-1%	1%
42	Hubbard County	20,428	0%	-1%	-1%	-1%	-1%	0%	0%	0%	-2%
43	Martin County	20,840	1%	-1%	0%	0%	0%	2%	1%	-2%	-1%
44	Fillmore County	20,866	0%	-1%	1%	1%	1%	-3%	1%	2%	-1%
45	Nobles County	21,378	0%	0%	-1%	-1%	0%	0%	2%	-2%	1%
46	Wabasha County	21,676	1%	0%	1%	1%	1%	-2%	0%	3%	-1%
47	Meeker County	23,300	1%	0%	0%	0%	1%	0%	-1%	0%	1%
48	Todd County	24,895	-1%	1%	-2%	1%	0%	-1%	0%	0%	0%
49	Lyon County	25,857	1%	1%	-1%	1%	0%	-1%	-1%	1%	1%
50	Brown County	25,893	0%	1%	-1%	0%	0%	1%	0%	0%	-1%
51	Mille Lacs County	26,097	0%	2%	-1%	0%	-1%	1%	-3%	3%	0%
52	Le Sueur County	27,703	1%	0%	-1%	0%	1%	0%	-3%	1%	0%
53	Cass County	28,567	0%	0%	0%	0%	0%	0%	2%	-6%	0%
54	Pine County	29,750	-1%	0%	1%	1%	-1%	-1%	0%	1%	-1%
55	Freeborn County	31,255	0%	0%	0%	0%	0%	0%	1%	1%	1%
56	Polk County	31,600	0%	-1%	1%	-1%	1%	2%	1%	-2%	-1%
57	Becker County	32,504	-2%	1%	-1%	-1%	1%	-2%	0%	2%	0%
58	Nicollet County	32,727	-1%	-1%	1%	0%	0%	-2%	3%	-1%	0%
59	Morrison County	33,198	0%	0%	0%	0%	0%	1%	0%	-1%	0%
60	Carlton County	35,386	0%	-1%	2%	0%	2%	2%	-3%	1%	1%
61	Douglas County	36,009	-1%	0%	0%	1%	0%	1%	-1%	0%	0%
62	Steele County	36,576	0%	-1%	0%	0%	-1%	2%	-1%	-2%	-1%
63	McLeod County	36,651	0%	0%	0%	0%	0%	1%	-1%	2%	1%
64	Isanti County	37,816	0%	0%	1%	0%	0%	0%	-2%	0%	0%
65	Benton County	38,451	-1%	1%	-1%	0%	1%	0%	-1%	1%	0%
66	Mower County	39,163	1%	1%	-1%	-1%	0%	1%	0%	0%	0%
67	Kandiyohi County	42,239	-1%	0%	0%	0%	-1%	-1%	2%	0%	0%
68	Beltrami County	44,442	1%	0%	0%	-1%	-1%	1%	1%	2%	0%
69	Itasca County	45,058	0%	0%	0%	0%	0%	-1%	1%	2%	2%
70	Goodhue County	46,183	1%	0%	0%	0%	0%	1%	0%	-1%	0%
71	Winona County	51,461	1%	0%	1%	0%	0%	-1%	2%	0%	1%
72	Chisago County	53,887	0%	1%	-1%	0%	0%	-1%	0%	3%	-1%
73	Otter Tail County	57,303	0%	0%	0%	0%	0%	0%	0%	1%	0%
74	Clay County	58,999	0%	-1%	1%	1%	0%	-1%	1%	0%	0%
75	Crow Wing County	62,500	0%	0%	0%	0%	0%	0%	1%	0%	-1%
76	Blue Earth County	64,013	0%	1%	-1%	0%	0%	-1%	2%	-1%	0%
77	Rice County	64,142	0%	-1%	1%	0%	0%	-2%	1%	0%	1%
78	Sherburne County	88,499	0%	0%	0%	0%	1%	-1%	-1%	1%	0%
79	Carver County	91,042	0%	0%	0%	0%	0%	1%	-1%	1%	0%
80	Wright County	124,700	0%	0%	0%	0%	0%	0%	0%	0%	0%
81	Scott County	129,928	0%	-1%	1%	0%	0%	1%	0%	0%	0%
82	Olmsted County	144,248	0%	0%	0%	0%	0%	1%	-1%	0%	0%
83	Stearns County	150,642	0%	0%	0%	0%	0%	-1%	1%	0%	0%
84	St. Louis County	200,226	0%	0%	0%	0%	0%	0%	0%	0%	0%
85	Washington County	238,136	0%	0%	0%	0%	0%	-1%	1%	1%	0%
86	Anoka County	330,844	0%	0%	0%	0%	0%	0%	0%	0%	0%
87	Dakota County	398,552	0%	0%	0%	0%	0%	0%	0%	0%	0%
88	Ramsey County	508,640	0%	0%	0%	0%	0%	0%	0%	0%	0%
89	Hennepin County	1,152,425	0%	0%	0%	0%	0%	0%	0%	0%	0%
90	Minnesota	5,303,925	0%	0%	0%	0%	0%	0%	0%	0%	0%

12. Walter Schwarm, California Dept. of Labor

Below are some comments on an analysis of the 1st publicly released DHC demonstration file. Thank you for the opportunity to examine and provide input on these critical data.

- Differences in median age for off-spine geographies such as Places including CDPs are much bigger than for the on-spine geographies. Places and CDPs need more accurate age pyramids if regional planning, projections, and analysis is to be successfully performed.
- Severing the relationship between persons and households is the greatest source of problems in the data. Data users have specific long-standing needs for reasonable data in this area including:
  - Persons per Household (household population divided by number of households) which is used in the Housing Unit method for estimation.
  - Headship rates (householders of a certain age divided by household population of that age) are used project future housing needs as well as examining the age structure of housing.
  - Home ownership by ethnicity (householders by race/ethnicity divided by population of that group) which is a major component of equity and access policy.

    Inconsistencies in the data require data users to use alternative data sources and methods that are prone to high error rates. Basic information on the overall composition of neighborhoods and small places is rendered virtually useless as a result of the inconsistencies. Reasonable policy and planning can not be performed at the geographical level that is required.

- Table cells for household type and household size that generally have lower counts (less common household types and household sizes) often have very large percentage errors (e.g. over 30% of tracts have more than 10% error) which severely limits the usability of these tables.
- The shift of Housing units in Places between Occupied and Vacant is troubling as it directly affects popular small area estimation techniques such as the Housing Unit Method.

CA- Number of Housing Units	**Occupied Housing Units**		**Vacant Housing Units**	
	SF1	**DHC**	**SF1**	**DHC**
Less than 25	2,385,889	2,306,016	245,257	325,130
Between 25 and 49	3,012,390	3,013,637	224,857	223,610
Between 50 and 99	2,710,016	2,739,115	212,420	183,321
Between 100 and 249	2,768,995	2,805,429	244,874	208,440
Between 250 and 499	1,249,097	1,260,109	121,737	110,725
Between 500 and 999	414,128	416,225	47,390	45,293
Equal to or Greater than 1,000	36,983	37,004	6,048	6,027

- Group Quarters, particularly on the Non-Institutional side show shifting population levels that would significantly change analysis and programmatic results were they used

California	Number of places with fewer persons under DHC	Maximum Population Reduction	Median Percent Reduction	Total Places
Institutional GQ	235	(83)	(4.3)	544
Correctional Facilities	71	(74)	(1.5)	155
Non-Institutional GQ	304	(222)	(9.4)	820

Of further concern are the number of blocks that have fatal inconsistencies which would invalidate any attempt to use the data to build either a custom geography (say a special district) or in tracking disasters, or public health issues:

CA-DHC blocks with population but no SF1 population: 5,585;
CA-SF1 blocks with population but no DHC population: 2,650;

CA-SF1 blocks with 0-17 population but no 18+ population: 59; CA-DHC blocks with 0-17 population but no 18+ population: 3,848;

CA-Blocks where population flipped from majority minority to majority white: 10,692; CA-Blocks that flipped from white majority to non-white majority: 33,263;

CA-DHC blocks with household population but no occupied housing units: 21,016; CA-DHC blocks with occupied housing units but no household population: 1,795;

CA-DHC blocks with more than 15 persons per household: 4,298; CA-SF1 blocks with more than 15 persons per household: 16;

CA	Total Population		Persons Age 0-17		Persons Age 18+	
Person Count	MAE	MALPE	MAE	MALPE	MAE	MALPE
Less than 50 persons	5.76	57.738	3.020	50.716	4.032	36.696
Between 50 and 249 persons	7.70	(0.526)	4.705	1.544	5.561	(0.320)
Between 250 and 499 persons	13.01	(3.061)	7.259	(4.417)	8.179	(2.191)
Between 500 and 749 persons	19.36	(3.012)	10.086	(4.362)	11.344	(2.181)

Between 750 and 999 persons	24.328	(2.721)	12.449	(3.979)	13.999	(1.947)
More than 1,000 persons	30.540	(2.247)	15.438	(0.330)	17.225	(1.639)

Even when considering larger geographical areas, there are inconstancies that make building custom geographies potentially fraught with undesirable and unrealistic outcomes:

CA-DHC block groups with population but no SF1 population: 50;
CA-SF1 block groups with population but no DHC population: 50;

CA-Block groups where population flipped from majority minority to majority white: 90; CA-Block groups that flipped from white majority to non-white majority: 126;

CA-SF1 urban block groups switched to rural BG: 13; CA-SF1 rural BG switched to urban BG: 7;

13. Abraham D Flaxman, Institute for Health Metrics and Evaluation, University of Washington

Attached please find our investigation into a specific disclosure risk:

> We conducted a simulation study to investigate the risk of disclosing a change in how an individual's sex was recorded in successive censuses. In a simulated population based on a reconstruction of the 2010 decennial census of Texas, we compared the number of transgender individuals under 18 identified by linking simulated census data from 2010 and 2020 under alternative approaches to disclosure avoidance, including swapping in 2020 (as used in the 2010) and TDA in 2020 (as planned for the actual release).
>
> We found that without any disclosure avoidance in 2010 or 2020, a reconstruction-abetted linkage attack identified over 500 transgender children. With 5% swapping in 2010 and 2020, it identified 461 individuals, a 12% decrease. With swapping in 2010 and TopDown in 2020, it identified 61 individuals, an 88% decrease from swapping.

I hope you find this helpful. Thank you for your work!

THE RISK OF LINKED CENSUS DATA TO TRANSGENDER CHILDREN: A SIMULATION STUDY

ABRAHAM D. FLAXMAN AND OS KEYES

Institute for Health Metrics and Evaluation, University of Washington

e-mail address: abie@uw.edu

Abstract. Every ten years the United States Census Bureau collects data on all people living in the US, including information on age, sex, race, ethnicity, and household relation- ship. They are required by law to protect this data from disclosure where data provided by any individual can be identified, and, in 2020, they used a novel approach to meet this requirement, the differentially private TopDown Algorithm.

We conducted a simulation study to investigate the risk of disclosing a change in how an individual's sex was recorded in successive censuses. In a simulated population based on a reconstruction of the 2010 decennial census of Texas, we compared the number of transgender individuals under 18 identified by linking simulated census data from 2010 and 2020 under alternative approaches to disclosure avoidance, including swapping in 2020 (as used in the 2010) and TDA in 2020 (as planned for the actual release).

We found that without any disclosure avoidance in 2010 or 2020, a reconstruction-abetted linkage attack identified over 500 transgender children. With 5% swapping in 2010 and 2020, it identified 461 individuals, a 12% decrease. With swapping in 2010 and TopDown in 2020, it identified 61 individuals, an 88% decrease from swapping.

In light of recent laws prohibiting parents from obtaining medical care for their trans children, our results demonstrate the importance of disclosure avoidance for census data, and suggest that the TopDown approach planned by Census Bureau is a substantial improvement compared to the previous approach, but still risks disclosing sensitive information.

Introduction

As part of the 2020 decennial census, the US Census Bureau has developed a new approach to disclosure avoidance, based on differential privacy, called the TopDown Algorithm (TDA) (Abowd et al. [2019]). The details of their approach have been refined iteratively since they first debuted as part of the 2018 end-to-end test (Garfinkel et al. [2019]). The release of the Demographics and Housing Characteristics (DHC) data in August, 2023 will be the next application of TDA for a data product from the 2020 decennial census. At this time of writing (May 2022) we have the products of the first application of TDA (the Public Law 94-171 redistricting data, released in August, 2021) as well as a demonstration DHC product from a test run in March 2022 (Bureau [2022]) to help us understand plans and trade-offs for some of the TDA options previously enumerated (Petti and Flaxman [2019]).

Key words and phrases: census, disclosure avoidance, linkage attack.

In support of their work to develop and validate TDA, the Census Bureau has previously released a series of Privacy-Protected Microdata Files (PPMFs) by applying iterations of TDA to the 2010 Census Edited File. The DHC product from March 2022 diverges from this pattern and provides summary tables without releasing a corresponding PPMF. This invites the question of whether the release of a PPMF or reconstruction of microdata from DHC tables might compromise privacy. In this work, we investigated empirically how well TDA protects against disclosure of sensitive information on an individual's gender identity in DHC data.

Past investigations of demonstration products have focused primarily on the impact of TDA on accuracy of key census-derived statistics, and we agree that there are broad, political implications behind statistical accuracy; the framing of census data informs everything from the shape and number of legislative districts to funding and resourcing for minority groups (see, for example, Thompson [2012]). But this is also true of privacy—accurate representation is not an unalloyed good. For many groups, particularly those who are vulnerable to and have experienced active discrimination by state entities, higher accuracy can also mean higher identifiability and higher *scrutiny*. An example of this is undocumented immigrants' relation to questions about citizenship—questions that can be used to identify, surveil, and punish people who are undocumented, and consequently lead to reduced engagement with and trust of the census (see Barreto [2019]). More recent in the public eye (although just as longstanding, as highlighted by Canaday [2009]) are questions of gender (Singer [2015]), on which this investigation is focused.

The last few years have seen heightened scrutiny of transgender people (henceforth "trans"), with a particular focus on (and moral panic around) trans children (see Slothouber [2020]). This has included actions by state actors to simultaneously legislate against access to care and equal treatment, and use existing mechanisms of government to punish the children and parents who have become identifiable. Most prominently, the governor of Texas, in Abbott [2022], has directed the state Department of Family and Protective Services to investigate the parents of any trans child who receives gender-affirming medical care. In order to do so, he advocates drawing on existing systems for child and parent surveillance, including abuse reporting requirements, to identify targets.

As all of this suggests, there are many reasons for us to be cautious around data availability and the pursuit of accuracy as an untrammelled good. While it is beneficial from a statistical perspective, an absence of privacy simultaneously risks both producing real, material harms for the individuals identified, and undermining trust in the census itself and so (paradoxically) reducing the very accuracy that is aimed for. To demonstrate the importance of factoring identifiability into account—and the necessity of an emphasis on disclosure avoidance in census policy—we used simulation to investigate a risk to privacy, by focusing on the risk of disclosing a child's transgender status, through discordant reporting of binary gender in successive censuses.

Methods

We used computer simulation to compare the number of trans children who might be identified in a synthetic population under alternative scenarios of disclosure avoidance. Our approach began with a synthetic population of size and structure similar to the state of

Texas, derived from a reconstruction of the US population on April 1, 2010. Since our focus is on linking youth between the 2010 and 2020 Decennial Censuses, we included simulants from this population who were aged zero to seven and therefore would be under 18 on April 1, 2020. We augmented this reconstruction by assigning the simulant's gender based on responses to the Sexual Orientation and Gender Identity (SOGI) module of the Behavioral Risk Factors Surveillance System (BRFSS) collected in 2019 (National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health [2019]).

We initialized each simulant with attributes for age, gender, race, ethnicity, and household, where age was an integer value representing the age in years, gender was a five-valued variable (with values of transgender boy; transgender girl; transgender, gender nonconform- ing; cisgender boy; and cisgender girl), race was a 63-valued variable encoding the possible combinations of the six Census racial categories, ethnicity was a two-valued variable for Hispanic/non-Hispanic, and household was an identifier that encoded census geography (state, county, tract, block) as well as housing unit id.

From this initial population, we simulated the progression of time and the data captured in the 2010 and 2020 Decennial Censuses as follows: we recorded the age at initialization precisely for each simulant's reported age in the 2010 census, and then that age plus 10 for each simulant's reported age in the 2020 census. We used a simple model of the other key demographic factors of births, deaths, in-migration, and out-migration to simulate how this population might change over the next decade. Since our interest was in linking between censuses, we focused on migration, and posited that every household might move, making it harder to link. To realize this household mobility, we selected households to stay unmoved from 2010 to 2020 independently, with probability value $p_{\text{stay}} = 23\%$ derived from the American Communities Survey (we obtained this value by calculating the sample-weighted proportion of with-children households that had been in residence for at least 10 years in the 2020 5-year Public Use Microdata Sample). We updated the 2020 address of each non-staying household by selecting a new household for them to move to uniformly at random from all synthetic households in Texas that were occupied on Census Day 2010.

Finally, we simulated the reported value of sex on the 2010 and 2020 Decennial Census. Our model of reported sex started from the assumption—uncertain though it is—that, in the 2010 Census, nearly all of the transgender youth aged zero to seven had their sex reported based on their gender-assigned-at-birth. We then assumed that, for some of the simulants with transgender identities, this would lead to differing responses in the 2020 Census. Based on this premise, we simulated responses on the 2010 and 2020 census according to the following cases: for cisgender boy simulants, we recorded their sex as male in 2010 and 2020, and similarly for cisgender girl simulants we recorded female. For transgender boy simulants, we recorded their sex as female in 2010 and recorded their sex with a value chosen uniformly at random from the set male, female in 2020. Similarly for transgender girl simulants, we recorded their sex as male in 2010 and with a value of female in 2020 with probability 50%. For transgender, gender nonconforming simulants we recorded their sex as the same value in 2010 and 2020, with the value chosen uniformly at random from the set male, female .

We recorded race and ethnicity identically in 2010 and 2020, matching the value of the simulant's race and ethnicity attributes.

We compared four alternative scenarios of disclosure avoidance: (1) extreme disclosure where names were published, allowing even households that moved to be linked between censuses; (2) tables with no disclosure avoidance, where names were not published, but there was no effort to swap or otherwise perturb the data in published tables; (3) disclosure avoidance by swapping, where 5% of households were exchanged with another household to protect privacy; and (4) differentially private disclosure avoidance, where the new TDA approach was used to protect against disclosure in published tables. We now describe our method of quantifying how many transgender simulants would have their gender identity revealed in each of these scenarios.

Extreme disclosure (Scenario 1): In this scenario, we assumed that linking on name, age, race, and ethnicity would be able to identify nearly all simulants with discordantly reported values for sex in the 2010 and 2020 censuses. We therefore counted all simulants with differing values reported for sex in 2010 and 2020 to estimate the number of trans youth who would have their gender identity revealed if census microdata including names were released. We hypothesized that this would total in the thousands or perhaps even tens of thousands.

No disclosure avoidance (Scenario 2): In this scenario, we assumed that only simulants who had a unique combination of age, race, ethnicity, and geography were at risk of having their gender identity revealed by a reconstructed-abetted linkage attack. Furthermore, we assumed that individuals who moved between the 2010 and 2020 censuses would not have their transgender status revealed and even individuals who were exposed by a unique combination of attributes in 2010 and did not move by 2020 *might* not have their transgender status revealed, if in-migration to their census block resulted in them no longer having a unique combination of attributes in 2020. We therefore identified all simulants who did not move and had a unique combination of attributes in 2010 and also in 2020, and counted the simulants in this group with differing values reported for sex in 2010 and 2020. This constituted our estimate of the number of trans youth who would have their gender identity revealed by a reconstruction-abetted linkage attack if the tables used for reconstruction were published with no disclosure avoidance measures. We hypothesized that this would total in the hundreds.

Swapping for disclosure avoidance (Scenario 3): We approached this scenario similarly to Scenario 2, but instead of using each simulant's geography directly in the reconstruction-abetted linkage attack, we first chose a random subset of simulants to have their reported location swapped to somewhere other than their true location. We achieved this with a simple model analogous to the model of migration described above, where we selected some households to report in a location that is not their actual location independently, with probability $p_{\text{swap}} = 5\%$ (we chose this value as a modeling assumption broadly aligned with the publicly available information about the Census Bureau's approach to disclosure avoidance in the 2010 Decennial Census). For each of the selected households, we chose a reported location by selecting a household uniformly at random from all synthetic households in Texas on Census Day 2010.

We then identified all simulants who did not appear to have moved, according to their (possibly swapped) reported location in the 2010 and 2020 censuses, who had a unique combination of age, race, ethnicity, and geography attributes recorded in both censuses, and counted the simulants in this group with differing values reported for sex in 2010 and 2020. This constituted our estimate of the number of trans youth who would have their gender identity revealed by a reconstruction-abetted linkage attack if the tables used for reconstruction were protected by swapping. We hypothesized that this total would be five to 10% lower than the total from the no-disclosure-avoidance scenario, and therefore also reveal sensitive information about hundreds of trans youth.

TDA for disclosure avoidance (Scenario 4): Due to time constraints, we were not able to approach this scenario in a way as analogous to Scenarios 1-3 as we would have preferred. With more time or computer savvy, we would have run TDA ourselves on the synthetic data, after simulating forward ten years. Instead, we used the Census Bureau's DHC demonstration product to generate our estimate of the risk of a reconstruction-abetted linkage attack in this scenario, which is more complicated to explain than the previous three scenarios.

We began with a reconstruction exercise, to come up with a reconstructed microdata file (ReMF) consisting of a row for each reconstructed individual and columns for the attributes of age, sex, race, ethnicity, and geography that was consistent with the tables from the demonstration DHC product for individuals age zero to 17. We similarly generated an ReMF from the corresponding SF1 tables published as part of the 2010 Decennial Census. Instead of initializing our synthetic population in 2010 and simulating the progression of time, we initialized our synthetic population in 2020, based on the individuals aged 10 to 17 in the SF1 ReMF. We then simulated the *regression* of time, going backwards from 2020 to the 2010 Census Day, when each simulant would be 10 years younger. We applied our migration model to keep the location in 2010 identical to that in 2020 for only a random fraction simulants, governed again by the parameter p_{stay}.

As in the other scenarios, we endowed each simulant with a gender attribute, which we calibrated to match to measurements from the 2019 BRFSS SOGI module. However, in this scenario, we first set the reported sex in 2020 to match that in the SF1 ReMF, and then set the gender attribute and reported sex in 2010 conditional on the reported sex in 2020. This allowed us to use the demonstration DHC as our proxy for the privacy afforded by TDA in 2020 in our assessment of the number of trans youth who would have their gender identity revealed by a reconstruction-abetted linkage attack using data protected by swapping in 2010 and TDA in 2020.

To complete this approach, we identified all simulants who had a unique combination of age, race, ethnicity, and geography attributes recorded in 2010, and identified which of these simulants matched a unique individual aged 10 years older in the DHC ReMF. For each of these simulants, we then compared the reported sex in the 2010 census with the reported sex in the 2020 census. We counted how many of these links were for simulants who were trans youth. We hypothesized that this would be at least an order of magnitude smaller than the total from the swapping-for-disclosure-avoidance scenario.

Results

Our synthetic population included 25,145,561 individual simulants, matching exactly the 2010 population count for Texas. We focused on the simulants aged zero to seven on April 1, 2010, of which we had 3,095,857. Among these simulants, 0.53% were trans, with 0.18% trans boys, 0.23% trans girls, and 0.12% gender nonconforming. Over the ten years simulation the majority of households moved at least once, and only 23% of simulants resided in the same census block in 2010 and 2020.

We found that in our scenario with extreme disclosure, where individual-level data with linkable names was published (Scenario 1), linking between 2010 and 2020 census data to identify individuals with discordantly reported values for sex would identify over 6,000 trans kids, accounting for 38% of all trans kids in our simulated version of Texas.

In our scenario where tables like those in SF1 or DHC were published precisely as enumerated, without any disclosure avoidance measures applied (Scenario 2), we found that migration and non-uniqueness substantially reduced the number of trans kids who's gender identity was revealed. However, there were still 667,072 individuals who were uniquely identified by the age, race, ethnicity, and location in 2010 and 268,492 of them did not move and were still identified uniquely in 2020. In our simulation, a reconstruction-abetted linkage attack in this scenario still identified over 500 trans kids.

In our next scenario (Scenario 3), we added swapping-based disclosure avoidance to the tables in Scenario 2, and we found that with respect to a reconstructed-abetted linkage attack, swapping acted similarly to a small boost in migration for prevent identifying trans kids. At the 5% swapping level we used in Scenario 3, we found that a reconstruction-abetted linkage attack identified 461 trans kids, a 12% reduction from the number identified in Scenario 2.

Our final scenario is the closest we considered to the approach proposed by Census Bureau in the most recently released demonstration product. In this scenario, we considered protecting the tables released from the 2010 census with swapping and the tables from the 2020 census with TDA (Scenario 4). We found that this afforded substantially more protection than the other scenarios we considered. Because of the alternative route we took to constructing this scenario, we used a different initial population, starting with 3,009,117 simulants ages 10 to 17 on April 1, 2020. We found that TDA was successful in preventing the bulk of the identifications from Scenario 3; in our simulation, a reconstruction-abetted linkage attack identified only 61 trans kids when TDA was used for disclosure avoidance on the 2020 tables, an 88% reduction in the number identified when swapping was used in Scenario 3.

1. Discussion

Our simulation results demonstrate the magnitude of the threat that a linkage attack designed to identify trans kids might pose. Were Census Bureau to publish microdata on the 2010 and 2020 census (Scenario 1), it would likely identify the transgender status of over 6,000 trans kids in Texas. In the approach underlying the most recent demonstration data, on the other hand, a reconstruction-abetted linkage attack would likely identify the transgender status of only 61 trans kids in Texas. We hope that this convinces some readers of the importance of including disclosure avoidance in Decennial Censuses.

The bulk of previously published investigations into the quality of TDA demonstration products have compared with published results from the 2010 Census, and often reported differences. But in such comparisons there is an important limitation, because they compare the (published) results of swapping to the (demonstration) results of TDA applied to the unswapped data. Thus the conclusion of such a comparison is typically limited to proving that the noise introduced by TDA is different than the noise from swapping. This investigation turns this limitation into a strength, since a reconstruction-abetted linkage attack between 2010 and 2020 Decennial Censuses *will* be linking data that has been swapped with data that has been protected by TopDown. Modeling in a simulation framework like the approach developed here could potentially also be used in future investigations to more directly compare the noise introduced by swapping to the noise introduced by TDA.

Limitations: There are at least three simplifying assumptions in this simulation model that constitute limitations which might be the focus of future work. First, the migration model is quite simplistic, and it is likely that further investigation could more accurately incorporate determinants of migration; the probability that a households has stayed unmoved between decennial censuses is likely to vary by household income, for example, which is an attribute that we did not include in our simulation, but could potentially add. Second, our simple model of how sex was reported in 2020 census for trans kids could also be more complex, although it is less clear what sources of data could inform adding this complexity. Third, in this work we assumed that race and ethnicity were unchanged between 2010 and 2020 censuses, but it is likely that evolving conceptions of race and ethnicity have led to some recording of differing values for some individuals, and this would result in some reduction in the number of links in a linkage attack. We conjecture that none of these simplifying assumptions have substantially changed the number of trans kids identified in our scenarios, however.

As mentioned in the methods section, our approach to Scenario 4 is more complicated than we would have liked, and once we figure out how to run a DHC configuration for TDA, we can address this by pivoting to a more familiar simulation paradigm, where we initialize the simulation in 2010 and run time forwards. However, the approach we used in this work has a strength alluded to above, because it uses SF1 data that has been perturbed by the swapping approach actually employed by Census Bureau in the 2010 Decennial Census, the details of which are not publicly available.

We would also like to emphasise three limitations specific to our model of trans children. First, our assumption of a uniform probability of markers changing between census years is no doubt an overly-simplistic one; we would expect that, in practice, the likelihood of changes is variable depending on both the respondent family's context and the individual perspective of the child and their parent(s). Second, the limited range of sex options on the census means that many trans children whose identities fall outside a simplistic binary do not alter their census markers. Third, we would expect differences in the amount of geographic mobility and consistency in household structure for trans families writ large, with one response to increasing scrutiny, at least for those with means, being to purposefully move their household. These limitations suggest this is in fact the *minimal* count of trans people identifiable through the current census approach to data disclosure, and that without changes to the data disclosure approach, well-intended efforts to increase the ability of Census Bureau instruments to record and represent trans people (see White House [2022]) could increase the risk of identifiability and harm.

Although the focus of this piece is on trans *children*—specifically, those under 18 in both the 2010 and 2020 census, with different sex records in each—it is worth emphasising that they are not the only people at risk. With the addition of more census tranches (say, 2000, or, going forward, 2030), the range of people at risk of disclosing their transgender status would expand to include trans adults, many of whom, if they have children, are also being targeted for additional scrutiny by state bodies.

Due to data limitations, we had to use computer simulation to conduct this investigation, but it would be possible for Census Bureau to replicate and expand on analyses such as this one internally, where they can use private data such as the Census Edited File, which is not available to outside researchers. The Census Bureau could reproduce this analysis using its internal unprotected data to understand how its implementation differs from this model. We encourage them to share with us how much this risk differs in the true implementation from the risk as modeled in this simulation.

We have made a replication archive of this work available online: https://github.com/aflaxman/linked_census_disclosure

Acknowledgment

The authors wish to acknowledge fruitful discussions with danah boyd.

References

G. Abbott. Letter to Masters, 2022. URL https://www.documentcloud.org/documents/21272649-abbott-letter-to-masters.

J. Abowd, D. Kifer, B. Moran, R. Ashmead, P. Leclerc, W. Sexton, S. Garfinkel, and A. Machanavajjhala. Census TopDown: Differentially private data, incremental schemas, and consistency with public knowledge. Technical report, U.S. Census Bureau, 2019.

M. A. Barreto. Expert Testimony: *NYAG New York v. U.S. Immigration and Customs Enforcement*, 2019. URL http://mattbarreto.com/papers/Declaration_of_Matthew_A._Barreto_-_NY.pdf.

U. C. Bureau. 2010 demonstration data for the Demographic and Housing Characteristics file (DHC) (v. 2022-03-16). Technical report, 2022. URL https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/02-Demographic_and_Housing_Characteristics/2022-03-16_Summary_File/.

M. Canaday. The straight state. In *The Straight State*. Princeton University Press, 2009.

S. Garfinkel et al. 2018 end-to-end test disclosure avoidance system design specification. Technical report, U.S. Census Bureau, 2019.

National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health. 2019 brfss survey data and documentation. Technical report, 2019. URL https://www.cdc.gov/brfss/annual_data/annual_2019.html.

S. Petti and A. Flaxman. Differential privacy in the 2020 US census: What will it do? Quantifying the accuracy/privacy tradeoff. *Gates Open Research*, 3, 2019.

T. B. Singer. The profusion of things: the "transgender matrix" and demographic imaginaries in US public health. *Transgender Studies Quarterly*, 2(1):58–76, 2015.

V. Slothouber. (de) trans visibility: moral panic in mainstream media reports on de/retransition. *European Journal of English Studies*, 24(1):89–99, 2020.

D. Thompson. Making (mixed-) race: census politics and the emergence of multiracial multiculturalism in the United States, Great Britain and Canada. *Ethnic and Racial Studies*, 35(8):1409–1426, 2012.

White House. FACT SHEET: Biden-Harris administration advances equality and visibility for transgender Americans, 2022.

14. Jill Kaneff, Northern Virginia Regional Commission

Dear U.S. Census Bureau Feedback Recipients,

On behalf of the Northern Virginia Regional Commission, I am submitting feedback on the April 2021 demonstration data product. Attached please find our memo, as well as our former memo submitted in December 2019 that was in response to the October 2019 demonstration data. Thanks for the opportunity to provide feedback.

Sincerely,
Jill Kaneff

Sr. Regional Demographer/GIS Analyst
Northern Virginia Regional Commission

nvrc
Northern Virginia Regional Commission

3040 Williams Drive | Suite 200 | Fairfax, VA 22031
p: 703-642-0700 | **f:** 703-642-5077
www.novaregion.org

May 17, 2022

To: U.S. Census Bureau
4600 Silver Hill Rd.
Washington, DC 20233

From: Robert W. Lazaro, Jr., Executive Director
Jill Kaneff, Senior Regional Demographer

Re: Feedback on U.S. Census Bureau Demographic and Housing Characteristics Demonstration Data Product First Round

Dear U.S. Census Bureau Feedback Recipients,

The Northern Virginia Regional Commission (NVRC) is a regional government agency and council in the Northern Virginia suburbs of Washington DC. NVRC is one of 21 regional councils and planning districts in the Commonwealth of Virginia. The NVRC regional council has thirteen member local governments. NVRC represents an area with a population of 2.5 million.

This memo is intended to provide feedback to the U.S. Census Bureau on the 2010 Demonstration Data Product – Demographic and Housing Characteristics (DHC) that the Census Bureau is using as a prototype to develop the 2020 Census DHC data products. We appreciate the Census Bureau's willingness to take feedback from data users.

Staff only had a couple of days in their full schedule to dedicate to review. To the extent possible, Northern Virginia Regional Commission reviewed a select set of commonly used local government variables. Block group and tract level reviews were performed for the following variables.

1. Age: Grouped the age categories in the groups Under 18, 18 to 64, and 65 and Over.
2. Household Tenure overall
3. Household Tenure of Hispanic Householders
4. Household Tenure of White Non-Hispanic Householders

These variables were compared to the original 2010 census block group and tract level data. The share percentages were calculated for comparative purposes. There were two substantive findings with the share calculations.

1) Those shares that were calculated with the denominator being the overall total population or overall total households (e.g., renter share of total households) produced results that were believed to be within reasonable margins of error for nearly all census block groups and tracts, except for a few areas that have small populations. Geographic areas with small populations, such as the census tract encompassing Dulles Airport, are clearly understood to have high margins of error and would not be usable because the Census Bureau has stated on many different occasions that areas of small population are not fit-for-use for analysis and instead should be aggregated with other geographies to form a larger area of analysis.

2) Those shares that were calculated with the denominator being a subgroup of the population or households produced a significant amount of unreliable and unusable results at the tract level. For example, the share of Hispanic households that are renter versus owner is calculated with Hispanic households as the denominator. In Northern Virginia there are 520 census tracts, and 72 (or 14%) had an absolute difference in Hispanic owner household share of greater than or equal to 10% when compared to the original 2010 Census. An acceptable absolute difference we believe would be less than 5%. An example of this issue is seen with Tract 51510200801. This tract has 901 total households with Differential Privacy (DP) and 896 with the original Census. This same tract had 76 Hispanic households, of which 38 were owner households with DP, while the original Census had 73 Hispanic households, of which 21 were owner households. These figures result in the share of Hispanic households that are owned being 50% with DP versus 29% with the original Census. This level of error makes this data very inaccurate and unusable for government policy making and service planning efforts. Governments need information like this to identify inequities. For all variables, sub-group share calculations will need to have margins of error that are similar to the margins of error of the shares of total households and total population in order to be considered accurate enough for use in planning and analysis.

One other thing, we would like to point out is that NVRC and our partners only had a couple of days of staff time to dedicate to review. A more extensive review would have been possible in these couple of days had we known that IPUMS NHGIS had prepared a demonstration data product that had processed all the Census Bureau's demonstration data files and merged this data with the original 2010 Census data. The Census Bureau's March 16 and April 14 news releases did not convey this. Instead, the newsletters took readers to the Census Bureau's FTP data files. In the next demonstration data round's newsletter, it would help the user community tremendously, if the availability of NHGIS data files were communicated. NVRC staff only learned about the NHGIS products on the Census Bureau's feedback deadline date.

We thank you for providing the opportunity to review the DHC demonstration data and provide feedback.

Sincerely,

Robert W Lazaro Jr

Robert W. Lazaro, Jr.

Executive Director

Jill Kaneff

Jill Kaneff

Regional Demographer